

JCTC

Journal of Chemical Theory and Computation

Ab Initio Protein Folding Using a Cooperative Swarm of Molecular Dynamics Trajectories

Neil J. Bruce and Richard A. Bryce*

School of Pharmacy and Pharmaceutical Sciences,
University of Manchester, Oxford Road, Manchester,
M13 9PT, United Kingdom

Received February 1, 2010

Abstract: The use of atomistic simulation techniques to directly resolve the protein tertiary structure from the primary amino acid sequence is hindered by the rough topology of the protein free energy surface and the resulting simulation time scales required. We explore here the use of a molecular dynamics technique based on swarm intelligence to identify the native states of two peptides and a Trp-cage miniprotein. In all cases, the presence of cooperative swarm interactions significantly enhanced the efficiency of molecular dynamics simulations in predicting the native conformation.

1. Introduction

Despite the vast conformational space available to them, globular proteins are able to fold rapidly to their unique native geometries.¹ Accurate prediction of these folded states, using only primary sequence information, is an important goal in pharmaceutical science, as knowledge of the structure of protein targets is an important step in rational drug design. Knowledge-based² and coarse-graining methods³ have had some success in predicting protein structure. There are associated drawbacks, however, for example, a dependence on available geometrically similar structures in the former approach and a lack atomic resolution for the latter, which may lead to omission of important structural features.

Atomistic simulations are ideally placed to provide direct molecular level insights into the structure and dynamics of proteins. However, such approaches are hindered by the simulation time scales required to observe folding events. The free energy surface upon which protein simulations operate is rugged and characterized by a broad range of barriers, at scales both lower and higher than that of thermal energy kT . Advanced

simulation methods^{4–8} seek to increase the rate at which these barriers are traversed, while maintaining the representative features of the free energy surface; these methods offer the potential to increase the rate at which folding events occur during simulation, allowing the study of larger systems with more complex folding mechanisms.

One route to enhanced sampling that has proved successful in other areas of computational chemistry involves artificial intelligence methods. Genetic algorithms have had a major impact as conformational search tools in protein–ligand docking.^{9,10} An alternative class of artificial intelligence methodology is based on *swarm intelligence*, the emergent behavior observed in nature when social animals, such as swarming insects, flocking birds, or schooling fish, act together cooperatively. In groups, the animals are able to show a greater searching efficiency than they would when acting individually. A concept originally applied computationally in 1989 to cellular robotic systems,¹¹ the swarm behavior is modeled by a set of simple rules which describe how individual agents act cooperatively within the system.

A recent swarm intelligence approach based on the behavior of ant colonies has proved successful in guiding molecular docking¹² and loop refinement.¹³ Another swarm intelligence approach, particle swarm optimization,¹⁴ has been used in the development of QSAR models¹⁵ and molecular docking.^{16,17} For the latter, the swarm algorithm exhibited significant improvements in the RMSD of pose for 37 protein–ligand complexes, when compared to GOLD (Darwinian genetic algorithm), AutoDock (Larmakian genetic algorithm), and the commonly used FlexX and DOCK methods.¹⁶ The swarm intelligence approach has been extended to molecular dynamics simulations using model potentials¹⁸ but to our knowledge has yet to be applied to protein folding problems. In this letter, therefore, we explore the use of a swarm algorithm to guide protein structure prediction via multicopy molecular dynamics (MD). The method is applied to two peptides and a Trp-cage miniprotein.

2. Methodology

Our approach follows the SWARM-MD method of Huber and van Gunsteren,¹⁸ using a swarm of replica simulations that interact to cooperatively search phase space. This cooperation occurs through the incorporation of a swarm potential into the dynamics of each replica, which acts to drive each swarm member toward the mean trajectory of the entire swarm. The swarm potential is given as

$$V(\{\phi^j\})_{\text{swarm}} = \sum_{j=1}^M A \exp[-Bd_{\text{rms}}(\phi^j)] \quad (1)$$

* Corresponding author tel.: (0)161-275-8345, fax: (0)161-275-2481, e-mail: R.A.Bryce@manchester.ac.uk.

where $d_{\text{rms}}(\phi^j)$ is the root-mean-square distance of swarm member j from the average location of the swarm (in dihedral space), given by

$$d_{\text{rms}}(\phi^j) = \left(\frac{1}{N} \sum_{i=1}^N (\phi_i^j - \langle \phi_i \rangle_{\text{swarm}})^2 \right)^{1/2} \quad (2)$$

and ϕ_i^j is dihedral angle i in swarm member j , $\langle \phi_i \rangle_{\text{swarm}}$ is the average value of dihedral angle i over M members of the swarm, and N is the number of dihedral angles used in calculation of $d_{\text{rms}}(\phi^j)$. A and B are parameters that we take to be the same for each swarm member. A defines the maximum strength of the swarm potential, while B defines the range over which it acts. The total dynamics of the system are therefore described by two sets of parameters: the social parameters, described by eq 1 above, which result in the cooperation between members of the swarm, and the cognitive parameters that describe the interatomic potentials of the molecular mechanics force field, which act independently on each swarm member.

We have implemented the SWARM-MD algorithm into a modified version of Amber 9.¹⁹ Details of implementation (in particular, our treatment of $\langle \phi_i \rangle_{\text{swarm}}$) and associated parameters are given in the Supporting Information. Following Huber and van Gunsteren,¹⁸ we take range B to be 0.8 rad^{-1} and explore specific values of strength A , as discussed below. On the basis of trial calculations for AEK17 in an implicit solvent, we find a swarm size of 16–20 as optimal (Supporting Information); therefore, in this study, we employ 20-replica swarms. The method is applied to the simulated annealing of polyalanine, AEK17, and Trp-cage. Computational details of these annealing simulations, including initial structure generation and subsequent secondary structure analysis, are also provided in the Supporting Information.

3. Results and Discussion

We evaluate the ability of swarm-enhanced MD to fold two small peptides, polyalanine [Ac-(Ala)₁₁-NH₂] and AEK17 [Ac-Ala-(Glu-Ala-Ala-Ala-Lys)₃-Ala-NH₂], into their known α -helical structures.^{20,21} We also consider the folding of the Trp-cage miniprotein in an implicit solvent. In each of the three systems, we compare the results of a 20-replica swarm-enhanced MD simulation to those of 20 independent MD simulations. We consider each test case in turn.

3.1. Polyalanine. Starting from extended conformations, simulated annealing of gas-phase polyalanine was performed over 1.2 ns. In the absence of a swarm potential ($A = 0.0 \text{ kcal/mol}$) during annealing, most of the 20 independent polyalanine replicas become kinetically trapped in nonhelical conformations, with only five replicas reaching a fully helical conformation. This results in a final average helicity of 29% (Figure 1a). Note that polyalanine simulation replicas are described as folding to a completely helical conformation if all nonterminal amino acid residues are assigned as helical in the final annealed conformation (equivalent to a helical content greater than 82%).

Using a swarm potential of strength $A = -50.0 \text{ kcal/mol}$, folding performance is significantly improved: the final average helicity across polyalanine replicas is 75% (Figure 1a). Fifteen of the 20 swarm members anneal to a fully helical structure by the end of the simulation. A stronger swarm potential ($A = -100$

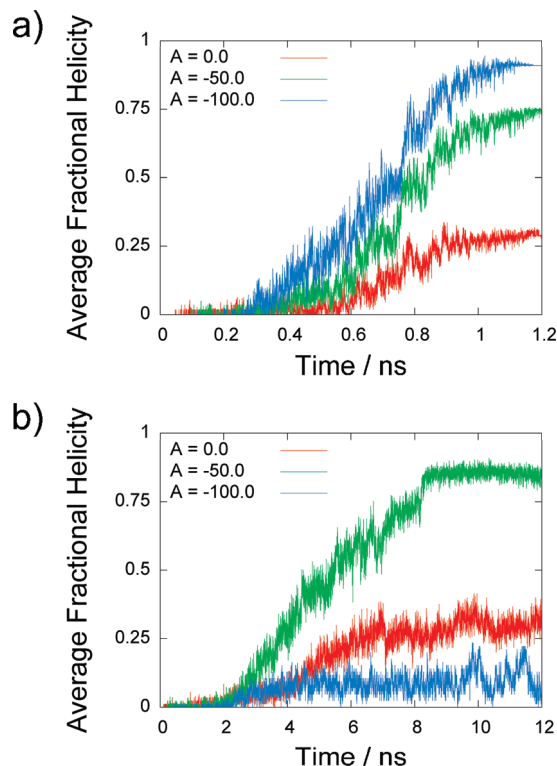


Figure 1. Average fractional helicity across 20 simulation replicas for swarm potentials of strength $A = -50.0$ (green) and -100.0 kcal/mol (blue) and in the absence of the swarm potential ($A = 0.0 \text{ kcal/mol}$, red) during simulated annealing of (a) gas-phase polyalanine and (b) AEK17 in aqueous solution.

kcal/mol) was also applied in a separate multicopy MD simulation and found to further improve performance. In this case, all replicas adopted a fully helical conformation, resulting in a final average helicity of 91% (Figure 1a).

3.2. AEK17. Second, we assess the ability of swarm-enhanced MD to fold the 17-mer AEK17 peptide, with its zwitterionic Glu-Ala-Ala-Ala-Lys repeat sequence, into an α helix. The peptide was modeled in an aqueous generalized Born (GB) solvent,²² and simulated annealing from extended conformations was conducted over 12 ns. AEK17 replicas are described as having successfully folded if they exhibit greater than 80% average helicity over the final 2 ns of constant temperature dynamics.

With no swarm potential applied, the final 20 annealed structures of AEK17 have a final average helicity of 34% (Figure 1b); only one simulation replica reaches a folded helical conformation. In the presence of a swarm potential ($A = -50.0 \text{ kcal/mol}$), 17 swarm replicas fold to completely helical conformations, resulting in final average helicity of 82% (Figure 1b). Interestingly, this agrees well with the experimentally observed value of $\sim 80\%$, measured by circular dichroism spectroscopy at 274 K and pH 7 in 0.01 M NaCl.²¹

A stronger swarm potential was also applied ($A = -100 \text{ kcal/mol}$), but found in this case to produce a negative effect on the observed folding rates. None of the simulation replicas was able to fold to a helical conformation during the 12 ns of simulation, resulting in an average final helicity of only 9% (Figure 1b). One possible explanation for the difference between this result and that of polyalanine, where the stronger potential was more

effective, may be found in the differing nature of the vacuum and aqueous potential energy surfaces. In a vacuum, unshielded interatomic Coulombic forces may be experienced at higher temperatures during annealing, leading to early prefolded nucleation steps in polyalanine. In contrast, in an aqueous environment, these forces are dampened; this is evidenced in independent MD simulations, where (un)folding transitions at ambient temperatures were found to occur more frequently in implicit aqueous solvent than *in vacuo*. As the temperature is lowered during annealing, the social swarm forces may dominate before cognitive MM forces in individual replicas sufficiently initiate physical folding processes. This results in the swarm force biasing the swarm members toward non-native high-energy minima, leading to premature convergence of the system. This observation underlines the importance of balancing social and cognitive potentials acting on the system.

3.3. Trp-cage. A more challenging model of protein folding is provided by the Trp-cage miniprotein (sequence: N₂₀LYIQWLKDGPPSSGRPPPS₃₉).²³ This small fast-folding (~4.1 μ s)²⁴ 20-residue protein forms a globular structure in solution and consists of an α helix (residues 20–28), a short 3₁₀ helix (residues 30–33), and a polyproline II helix (residues 36–38). The small size, and fast folding nature, of Trp-cage makes it an ideal test case to bridge the gap between studying small peptide systems and larger proteins. Indeed, it has already been the subject of much attention in protein folding and structure prediction studies, both through molecular dynamics simulation^{25–32} and nondynamical optimization.^{33–39}

Starting from extended conformations, a 20-replica 40 ns swarm-enhanced MD simulation of Trp-cage in GB solvent was performed. We adopt the strength of swarm potential ($A = -50$ kcal/mol) that gave the best folding performance for AEK17 in an implicit solvent. As with polyalanine and AEK17, we compare the performance of this simulation to that of 20 independent 40 ns molecular dynamics simulations.

For each simulation replica, the average root-mean-square deviation (RMSD) of both the backbone atoms and the heavy atoms, relative to the native NMR structure, was calculated, over the final 5 ns of simulation. In the absence of a swarm potential, the 20 independent replicas folded to conformations with an average backbone RMSD across all 20 simulations of 3.3 Å (Figure 2a) and an average heavy atom RMSD of 4.8 Å (Figure 2b). With the swarm potential applied, these values improved to 1.6 and 2.6 Å, respectively (Figure 2a and b). More specifically, of the 20 independent MD simulations, none was able to fold to within an average backbone RMSD of 1.5 Å (Figure 2a). In comparison, of the 20 swarm replicas, 16 folded to a backbone RMSD below 1.5 Å (Figure 2a). The lowest average backbone RMSD of any individual swarm member was 1.3 Å, significantly lower than the best performing independent replica, which had an average RMSD of 1.8 Å. The 16 folded swarm replicas all display an average heavy atom RMSD over the final 5 ns of below 2.3 Å (Figure 2b). Again, this is a significant improvement over nonswarm simulations, where the replica with the lowest average heavy atom RMSD over the final 5 ns differed from the NMR structure by 2.8 Å (Figure 2b).

The largest atomic deviations of the folded swarm members from the NMR structure of Trp-cage occur in its N- and

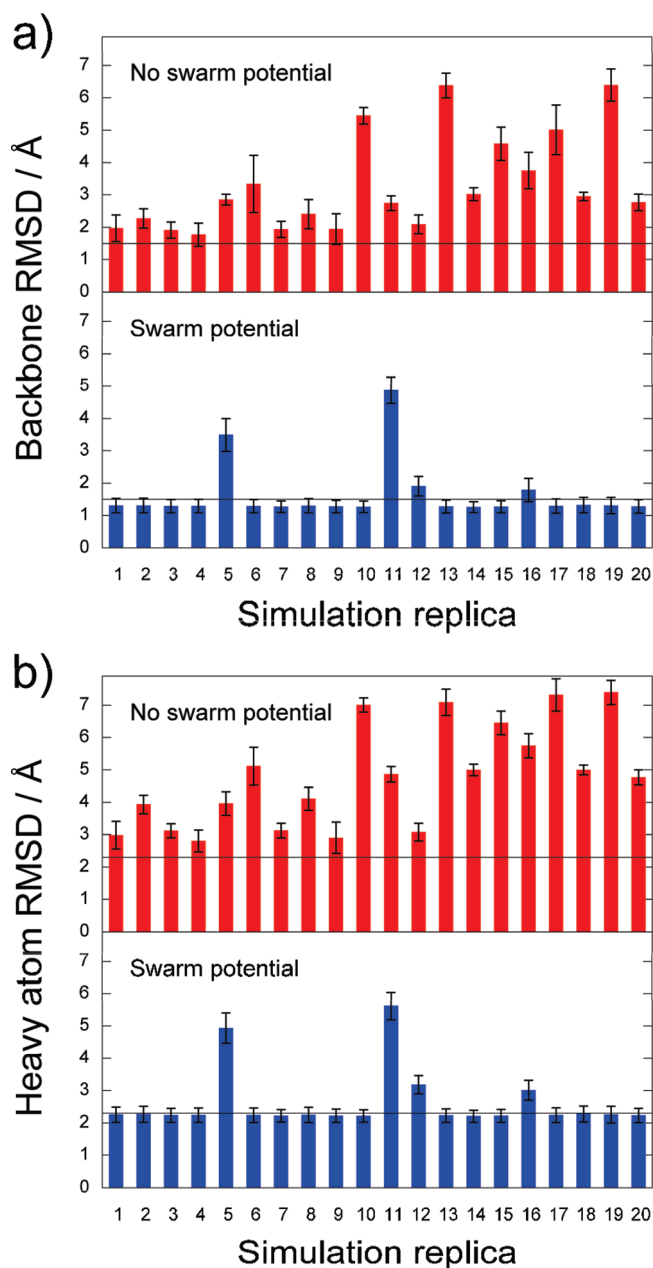


Figure 2. Mean RMS deviations of (a) backbone atom positions and (b) heavy atom positions from the native NMR structure of Trp-cage (conformer 11 from PDB code 1L2Y, which lies closest to average of NMR ensemble) during the final 5 ns of simulation for each simulation replica. Red indicates the absence of swarm potential; blue is with swarm potential present ($A = -50$ kcal/mol). Standard deviations during the final 5 ns of trajectories are shown as error bars.

C-terminal regions and the coil–3₁₀ helix–coil region (residues 29–35). These deviations are caused by the presence in the simulated structures of two salt bridges: between the terminal ammonium and carboxylate groups and between the side chains of Asp28 and Arg35. The terminal salt bridge is not present in any of the 38 NMR structures, while the Asp28–Arg35 interaction is present in approximately half the NMR structures. The persistence of these two salt bridges in the simulations results in the chain ends of Trp-cage lying closer in the folded structures compared to NMR (Figures 3a and 3b). The presence

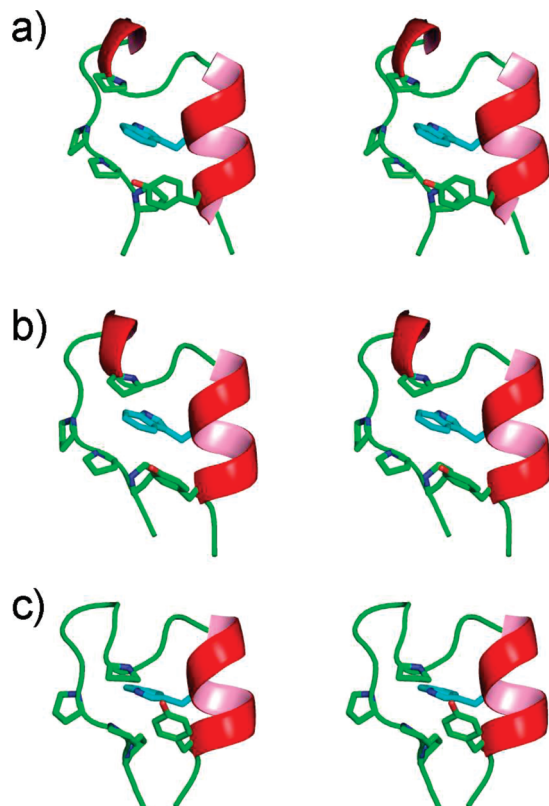


Figure 3. Stereoscopic image giving a comparison between the (a) Trp-cage NMR structure (PDB code 1L2Y; conformer 11), (b) best folded swarm member, and (c) a near-folded conformation with incorrect Trp25 orientation. The side chain of Trp25 is shown in blue; the side chains of hydrophobic cage residues Pro31, Pro36, Pro37, Pro38, and Tyr22 are shown in green.

of these interactions in 15 and 16 of the 16 folded swarm members suggests an overestimation of salt bridge strength, perhaps due to the force field–implicit solvent model combination (a known problem^{30,40}), which may be hindering further improvement in the heavy atom RMSD of the folded Trp-cage structures.

Four swarm members did not fold to within a backbone RMSD of 1.5 Å (Figure 2). Two swarm replicas of Trp-cage have a fairly high backbone RMSD (>3 Å) and contain only a partial native secondary structure: in both cases, the polyproline II helix was present. In one, the α helix was fully formed, while in the other, the α helix was partially formed but the 3_{10} helix was present. It is possible that these replicas represent intermediate structures in the folding mechanism of Trp-cage. The remaining two Trp-cage replicas have a backbone RMSD of less than 2.0 Å, corresponding to near-native states. However, the Trp25 side chain displays an incorrect orientation, pointing away from the cage (Figure 3c). It has been suggested^{25,26} that the rate-limiting step of Trp-cage folding is the incorporation of the Trp25 residue into the hydrophobic cage, followed by the formation of the residue's native contacts (Figure 3a). The resulting loss of degrees of freedom in this residue produces an entropic barrier in the free energy surface that must be overcome for the simulations to attain a native conformation.

It appears that the swarm-based simulations effectively lower this barrier: of the 18 swarm-enhanced Trp-cage replicas that

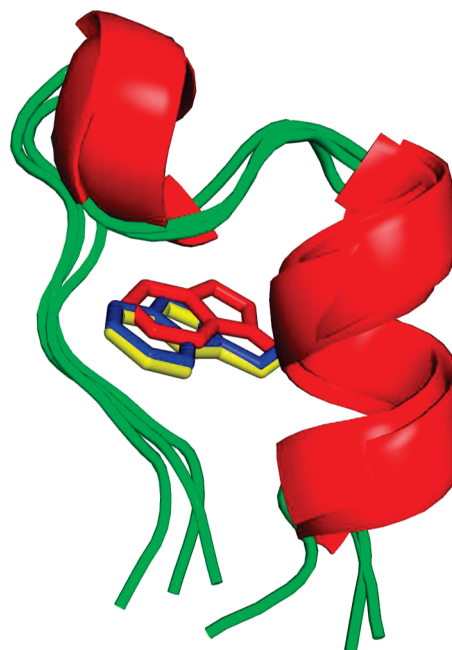


Figure 4. Overlay of NMR-derived Trp-cage and swarm-folded Trp-cage with the least NOE violations: Trp25 in NMR structure (blue) and best swarm structure before (red) and after refinement (yellow).

folded to a native backbone conformation (all with an average backbone RMSD < 2.0 Å), 16 (89%) displayed a correctly packed Trp25, pointing into the folded hydrophobic cage (Figure 3b), with just two showing the incorrect Trp25 geometry (Figure 3c). By contrast, only four of the nine (44%) independent MD replicas that produced a folded, or partially folded, hydrophobic cage (defined by an average backbone RMSD over residues 22–38 of less than 2.0 Å) displayed the correct Trp25 orientation. The addition of the cooperative swarm potential into the dynamics of the system therefore guides the simulations over the energy barrier associated with the entry of Trp25 into the cage, driving them toward the native state.

Interestingly, because of the swarm potential, the presence of a few replicas with mispacked Trp25 appears to slightly influence the orientation of that amino acid in the other replicas. While Trp25 packs correctly in these 16 replicas, the residue is subtly displaced relative to the NMR structure (Figure 4). This is clearly seen by direct comparison with the 168 NOE distance constraints used to derive the native structure ensemble of Trp-cage. In overall terms, the swarm-enhanced simulations violate only 25% of the NOE constraints, as compared to 34% from the independent simulations. However, the amino acid with the highest number of experimentally observed constraints (49 of the 168) is the core Trp25 residue. Consequently, because Trp25 tends to sit toward the back of the cage in the swarm, it is a replica from the independent MD, not swarm MD, that has the least NOE violations, with a value of 18% (replica 4, Figure 2). However, if the influence of the swarm on its replicas is tapered to zero in a subsequent refinement step, the Trp25 residue is able to assume its correct position, such that the best folded swarm member shows an NOE violation of only 14%. (This was achieved by annealing parameter *A* from -50.0 to

0.0 kcal/mol over 3 ns, followed by 2 ns of simulation in the absence of swarm potential.)

4. Conclusions

For polyalanine, AEK17, and Trp-cage, the incorporation of the swarm intelligence potential into the simulation dynamics is found to increase the folding performance of MD simulations. The cooperative nature of the swarm protocol prevents the swarm members from becoming trapped in local minima and helps drive the simulated structures toward the native state. Inhibition of AEK17 folding via the stronger swarm potential shows that a correct balance of social and cognitive factors is required for efficient conformational searching. However, this balance appeared suitably transferable between two implicitly solvated systems (AEK17 and Trp-cage).

For the study of Trp-cage folding, none of the 20 independent 40 ns molecular dynamics simulations anneal to a correctly folded conformation, i.e., to within a backbone RMSD of 1.5 Å of the native structure. By allowing the trajectories to interact cooperatively, 16 of the 20 trajectories adopt the native geometry within the last 5 ns of the simulations. A final refinement step, where the influence of the swarm potential was relaxed, was found to be useful in obtaining the detailed orientation of Trp25 in the miniprotein core. Both the swarm and independent MD calculations consume 800 ns of aggregate simulation time. To estimate the time scale required to correctly fold Trp-cage via unbiased molecular dynamics simulations at 300 K, we consider two recent generalized Born studies of the miniprotein: in the first study,²⁵ 77 independent 100 ns MD simulations of Trp-cage obtained only five conformations folded to within a backbone RMSD of 2.0 Å of the native state; a second MD study by Snow et al.³¹ harvested in the region of 1000 trajectories of length 30 ns, less than 1% of which folded to below 2.6 Å RMSD of the native C α structure. Simply considering computational cost per folded structure and recognizing that the SWARM-MD simulations also include the effect of annealing, the study of Snow et al. required 1–4 μ s per folded structure, compared to 50 ns per folded structure via the swarm-enhanced calculation, an improvement of 30- to 80-fold. In terms of overall computing time such that misfolded structures are included, the swarm simulations are 10 to 40 times shorter.

Other methods have been introduced to enhance the sampling of phase space by molecular simulation methods, such as metadynamics,⁴¹ locally enhanced sampling,⁴² and replica-exchange schemes.⁴ Several of these techniques have been applied to the folding of Trp-cage.^{27–29,32} Most recently, a temperature REMD simulation²⁷ was performed on Trp-cage in generalized Born solvent using 16 40 ns replicates spanning 300–460 K; the 300 K trajectory latterly sampled mainly folded Trp-cage (within \sim 2 Å heavy atom RMSD of the native structure). These calculation conditions and the method's performance are comparable to that of the SWARM-MD simulation of Trp-cage presented here. Hamiltonian-based replica exchange is also possible,^{27,28,32} and these methods have proved particularly powerful, for example, obtaining folded Trp-cage structures using five²⁷ or six³² replicas of sub-100 ns trajectories. An additional advantage of temperature-based and Hamiltonian-based replica exchange schemes is the generation of correct 300 K ensembles, providing information on folding

pathways and intermediates. However, neither of these methods scale favorably with system size due to the constraint of the exchange condition: efficient exchange between neighboring replicas requires sufficient energy overlap between replicas. Therefore, the number of required replicas grows rapidly with the size of the simulation system, and correspondingly longer simulation times are required to allow efficient sampling of the temperature space. SWARM-MD does not incorporate an exchange move and therefore does not suffer from this exacting requirement. This difference may be of greater importance when extending the method to study explicitly solvated systems, where system size is greatly increased by the degrees of freedom of water molecules. As mentioned previously with regard to salt bridge stability in Trp-cage, the absence of an explicitly modeled solvent can affect the free energy minima predicted through simulation, due to an incorrect representation of charge shielding.^{30,40} This overstabilization is not unique to Trp-cage.^{43,44} Existing implicit solvent models have also been found to incorrectly predict the secondary structure preferences of peptides⁴⁵ and overestimate melting temperatures.^{29,46} It has also been suggested that the absence of explicit water molecules may incorrectly describe the effects of dewetting of solvent-exposed hydrophobic residues in folded protein structures⁴⁷ and the effect of structural water molecules on the folding landscape.⁴⁸ From these observations, and the SWARM-MD algorithm's suitability for distribution of replicas over parallel architectures, a cooperative swarm of trajectories appears well-placed to predict the conformations of larger systems, particularly systems involving the incorporation of explicitly modeled bulk solvent. However, for future applications to predictive folding of larger polypeptide structures and interaction of flexible protein receptors with ligands, it will be useful to explore optimization schemes which anneal the influence of the swarm, to prevent the undue biasing effects of outlier replicas.

Acknowledgment. This work was supported by the EPSRC.

Supporting Information Available: Derivation of atomic forces due to swarm potential. Detailed simulation protocols. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Dill, K. A.; Chan, H. S. *Nat. Struct. Biol.* **1997**, *4*, 10–19.
- (2) Zhang, Y.; Skolnick, J. *Proc. Natl. Acad. Sci.* **2005**, *102*, 1029–1034.
- (3) Nania, M.; Chinchio, M.; Pillardy, J.; Ripoll, D. R.; Scheraga, H. A. *Proc. Natl. Acad. Sci.* **2003**, *100*, 1706–1710.
- (4) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (5) Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151–169.
- (6) Wu, X. W.; Brooks, B. R. *Biophys. J.* **2004**, *86*, 1946–1958.
- (7) Yang, L.; Grubb, M. P.; Gao, Y. Q. *J. Chem. Phys.* **2007**, *126*, 125102–125107.
- (8) Wolf, M. G.; de Leeuw, S. W. *Biophys. J.* **2008**, *94*, 3742–3747.
- (9) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727–748.

- (10) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
- (11) Beni, G. From swarm intelligence to swarm robotics. In *Swarm Robotics*; Sahin, E., Spears, W. M., Eds.; Springer-Verlag: Berlin, 2005; Vol. 3342, pp 1–9.
- (12) Korb, O.; Stutzle, T.; Exner, T. E. PLANTS: Application of ant colony optimization to structure-based drug design. In *Ant Colony Optimization and Swarm Intelligence, Proceedings*; Dorigo, M., Gambardella, L. M., Birattari, M., Martinoli, A., Stutzle, T., Eds.; Springer-Verlag: Berlin, 2006; Vol. 4150, pp 247–258.
- (13) Xiang, Z.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci.* **2002**, *99*, 7432–7437.
- (14) Kennedy, J.; Eberhart, R. Particle swarm optimization. In *1995 IEEE International Conference on Neural Networks Proceedings*; IEEE: New York, 1995; Vols 1–6, pp 1942–1948.
- (15) Cedeno, W.; Agrafiotis, D. K. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 255–263.
- (16) Chen, H. M.; Liu, B. F.; Huang, H. L.; Hwang, S. F.; Ho, S. Y. *J. Comput. Chem.* **2007**, *28*, 612–623.
- (17) Chen, K.; Li, T. H.; Cao, T. C. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 248–259.
- (18) Huber, T.; van Gunsteren, W. F. *J. Phys. Chem. A* **1998**, *102*, 5937–5943.
- (19) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M. *AMBER 9*, University of California: San Francisco, 2006.
- (20) Levy, Y.; Jortner, J.; Becker, O. M. *Proc. Natl. Acad. Sci.* **2001**, *98*, 2188–2193.
- (21) Marqusee, S.; Baldwin, R. L. *Proc. Natl. Acad. Sci.* **1987**, *84*, 8898–8902.
- (22) Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275–291.
- (23) Neidigh, J. W. *Nat. Struct. Biol.* **2002**, *9*, 425.
- (24) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952–12953.
- (25) Chowdhury, S.; Lee, M. C.; Duan, Y. *J. Phys. Chem. B* **2004**, *108*, 13855–13865.
- (26) Chowdhury, S.; Lee, M. C.; Xiong, G.; Duan, Y. *J. Mol. Biol.* **2003**, *327*, 711–717.
- (27) Kannan, S.; Zacharias, M. *Int. J. Mol. Sci.* **2009**, *10*, 1121–1137.
- (28) Piana, S.; Laio, A. *J. Phys. Chem. B* **2007**, *111*, 4553–4559.
- (29) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci.* **2003**, *100*, 7587–7592.
- (30) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258–11259.
- (31) Snow, C. D.; Zagrovic, B.; Pande, V. S. *J. Am. Chem. Soc.* **2002**, *124*, 14548–14549.
- (32) Son, W. J.; Jang, S.; Pak, Y.; Shin, S. *J. Chem. Phys.* **2007**, *126*, 5.
- (33) Carnevali, P.; Tóth, G.; Toubassi, G.; Meshkat, S. N. *J. Am. Chem. Soc.* **2003**, *125*, 14244–14245.
- (34) Schug, A.; Herges, T.; Verma, A.; Lee, K. H.; Wenzel, W. *ChemPhysChem* **2005**, *6*, 2640–2646.
- (35) Schug, A.; Herges, T.; Wenzel, W. *Phys. Rev. Lett.* **2003**, *91*, 158102.
- (36) Schug, A.; Wenzel, W. *Europhys. Lett.* **2004**, *67*, 307.
- (37) Schug, A.; Wenzel, W.; Hansmann, U. H. E. *J. Chem. Phys.* **2005**, *122*, 7.
- (38) Verma, A.; Schug, A.; Lee, K. H.; Wenzel, W. *J. Chem. Phys.* **2006**, *124*.
- (39) Verma, A.; Wenzel, W. *Biophys. J.* **2009**, *96*, 3483–3494.
- (40) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2005**, *2*, 115–127.
- (41) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562–12566.
- (42) Simmerling, C.; Fox, T.; Kollman, P. A. *J. Am. Chem. Soc.* **1998**, *120*, 5771–5782.
- (43) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci.* **2002**, *99*, 12777–12782.
- (44) Zhou, R. H. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 148–161.
- (45) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420–433.
- (46) Yeh, I. C.; Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2008**, *112*, 15064–15073.
- (47) Daidone, I.; Ulmschneider, M. B.; Di Nola, A.; Amadei, A.; Smith, J. C. *Proc. Natl. Acad. Sci.* **2007**, *104*, 15230–15235.
- (48) Rhee, Y. M.; Sorin, E. J.; Jayachandran, G.; Lindahl, E.; Pande, V. S. *Proc. Natl. Acad. Sci.* **2004**, *101*, 6456–6461.

JCTC

Journal of Chemical Theory and Computation

How Different are Electron-Rich and Electron-Deficient π Interactions?

Inacrist Geronimo, Eun Cheol Lee, N. Jiten Singh,* and Kwang S. Kim*

Center for Superfunctional Materials, Department of Chemistry, Pohang University of Science and Technology, Pohang, 790-784, Korea

Received April 5, 2010

Abstract: The intermolecular interaction driven structural change is vital to molecular architecturing. In the Cambridge Structural Database (CSD), we find that the preference for geometrical conformations of electron-deficient π systems is different from those of electron-rich π systems. Indeed, ab initio calculations find that electron-deficient π ring systems should involve different structures and energetics, consistent with the CSD search, due to the electric multipole moments and the decrease in the spatial extent of π -electron density.

Introduction

The rational design of nanomaterials relies on an understanding of noncovalent interactions, including hydrogen bonding and π interactions,¹ which enables the reversible self-assembly of supramolecular aggregates,² folding of proteins,³ and stacking of DNA.⁴ Most studies conventionally employ electron-rich π systems, such as the benzene dimer and mixed complexes with substituted benzene, as a model of aromatic π interactions.⁵ Electron-deficient systems should involve different energetics due to electric multipole moments and a decrease in the spatial extent of π -electron density.^{6–12} Quadrupole moments (Q_{zz} perpendicular to the ring plane) become more positive relative to benzene (Bz) in isoelectronic N-containing heterocycles pyridine (Py), pyrazine (Pz), triazine (Tr), and tetrazine (Tt) (−8.8 DÅ to 3.3 DÅ, Supporting Information). Therefore, it is of importance to investigate the changes of conformational preference against the number of N atoms in the ring ($\#_N$). The structural change is vital to designing intriguing structures for molecular architecture.¹³ Model systems that have been investigated include pyridine,^{7,8} pyrazine,⁹ pyrimidine,^{10,11} and triazine.^{11,12}

* Corresponding author e-mail: kim@postech.ac.kr (K. S. K.), jiten@postech.ac.kr (N. J. S.).

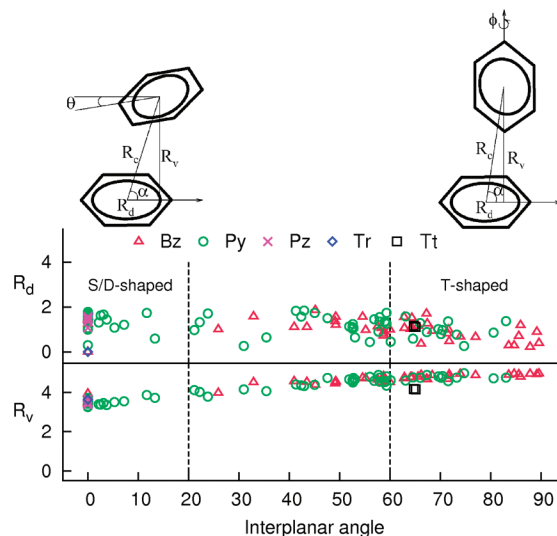


Figure 1. Scatterplots of the horizontal displacement (R_d) and vertical separation (R_v) vs the interplanar angle for benzene (Bz), pyridine (Py), pyrazine (Pz), triazine (Tr), and tetrazine (Tt) for their homopairing in the CSD. Interplanar angles less than 20° are labeled as stacked/displaced-stacked pairs, while those above 60° correspond to T-shaped pairs. The displacement angle is labeled as α . See the Supporting Information (Table S3) for the deviation from coplanarity θ (positive for upward tilt) in the displaced-stacked conformer and the rotation from the xz plane φ (counterclockwise direction taken as positive) in the T-shaped conformer.

Stacking interactions is a recurring motif in crystal structures containing heteroaromatic moieties, as demonstrated by crystal database searches on nitrogen-containing heterocyclic systems.¹⁴ A survey of unsubstituted fragments of N-containing heterocyclic pairs in organic crystals in the CSD indicates a preference for geometrical conformations different from that of benzene (see the Supporting Information for details). Figure 1 is a scatterplot of the horizontal displacement (R_d) and vertical separation (R_v) versus the interplanar angle.

Pairs of Bz moieties generally adopt a T-shaped orientation. Pairs of Py moieties show a wide distribution of displaced-stacked and T-shaped conformers. Pairs of Pz, Tr, and Tt moieties show displaced-stacked, stacked, and nearly T-shaped arrangements, respectively. Antiparallel orientations are observed for stacked and displaced-stacked Py and Tr moieties. T-shaped Py pairs involve C–H $\cdots\pi$ interactions, while Tt pairs have N $\cdots\pi$ (Supporting Information).

Calculations on the pyridine dimer^{7,8} yield a lower binding energy for the antiparallel conformer, and it was further shown by Piacenza and Grimme⁸ using density functional theory with

empirical dispersion corrections that the fully optimized structure is actually slightly bent from the perfect orientation at $\sim 160^\circ$. A cross-displaced stacked dimer was found to be the most stable geometry for pyrazine.⁹ The molecular electrostatic potential of triazine suggests favorable binding for the stacked structure for which the energy had been determined using rigid monomers at a vertical separation of 3.4 Å.¹¹ However, no direct comparison has been made with either fully optimized parallel¹² or antiparallel displaced-stacked conformers. The stacked conformer, for which theoretical studies are rather limited, is generally less stable and was shown to be a maximum in the potential energy curve for the displaced-stacked pyridine dimer.⁷ Detailed symmetry-adapted perturbation theory (SAPT) calculations have been limited to pyridine dimers. Preference for the antiparallel dipole orientation can be traced to electrostatic effects, while induction effects were reported to be an important stabilizing factor in T-shaped conformers. The potential energy curve for the displaced-stacked conformer also shows a decrease in the importance of exchange-repulsion contributions with increasing distance.⁷

Without an understanding of the interactions between electron-deficient π systems, one might still think that it is similar to previously studied π interactions of electron-rich centers. To clarify the issue, we carried out binding energy calculations for diverse electron-deficient π systems comprised of isoelectronic N-containing heterocyclic dimers at the complete basis set (CBS) limit using the coupled cluster theory with single, double, and perturbative triple excitations [CCSD(T)] and analyzed the energy components using SAPT. The systematic and accurate analysis enables us to distinguish the magnitude, directionality, and nature of interaction of electron-deficient π systems from those of well-known electron-rich π systems. To find the trend and origin governing such structural preferences, we investigate the impact of a progressive decrease in the π -electron density of the arene on both geometry and energy. Dimers of Bz ($\#_N = 0$), Py ($\#_N = 1$), Pz ($\#_N = 2$), 1,3,5-Tr ($\#_N = 3$), and 1,2,4,5-Tt ($\#_N = 4$) are used as prototypes.

Computational Method

The resolution of identity approximation of the second-order Møller–Plesset perturbation theory (RI-MP2)¹⁵ using the aug-cc-pVDZ (aVDZ) basis set with basis set superposition error (BSSE) correction was used to optimize the geometries of various parallel and T-shaped structures of pyridine, pyrazine, 1,3,5-triazine, and 1,2,4,5-tetrazine dimers. Single point energy calculations were subsequently performed at the RI-MP2/aug-cc-pVTZ (aVTZ) and CCSD(T)/aVDZ level with BSSE correction to obtain energies at the complete basis set (CBS) limit. The MP2 CBS limit was evaluated by using the extrapolation scheme based on the proportionality of the basis set error in the electron correlation energy to N^{-3} for the aug-cc-pVNZ basis set.¹⁶ This was then used to calculate the CCSD(T)/CBS limit. The total interaction energy was decomposed into electrostatic (E_{es}), induction (E_{in}), dispersion (E_{dp}), and exchange-repulsion (E_x) components based on SAPT.¹⁷ Here, E_{in} and E_{dp} includes the exchange-induction term and exchange-dispersion term, respectively, while E_x excludes the aforementioned terms from the exchange term. E_{dp} also includes the correction of the dispersion energy based on the difference between the CCSD(T)/

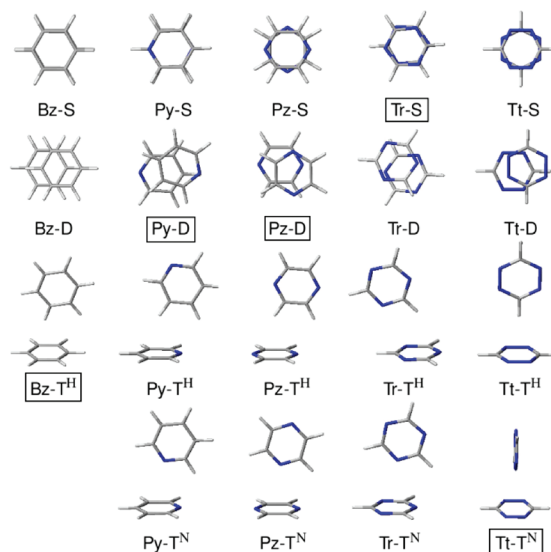


Figure 2. Basis-set-superposition-error (BSSE)-corrected RI-MP2/aug-cc-pVDZ optimized geometries of dimers of Bz, Py, Pz, Tr, and Tt. The lowest energy dimers are enclosed in boxes.

CBS and MP2/aVDZ binding energies (Supporting Information). The CCSD(T)/CBS total energy was added to correct the basis set dependency of the dispersion energy. SAPT calculations were performed with SAPT2008¹⁸ at the MP2/aVDZ' level where the p diffuse functions on H and the d diffuse functions on heavy atoms are removed. RI-MP2 and CCSD(T) calculations were done by using Turbomole 6.0.2¹⁹ and Molpro 2006.1,²⁰ respectively.

Results and Discussion

Stacked (S), displaced-stacked (D), and T-shaped (T^H , $H \cdots \pi$ interaction; T^N , $N \cdots \pi$ interaction) isomers of the Bz, Py, Pz, Tr, and Tt dimers are illustrated in Figure 2. To obtain large electrostatic interactions (either dipole–dipole or quadrupole–quadrupole interactions), the stacked or displaced-stacked Py and Tr dimers show antiparallel orientations between two monomers, while the stacked or displaced-stacked Pz and Tt dimers show perpendicularly rotated orientations. Related binding energies at RI-MP2 and CCSD(T)/CBS, along with the energy components based on SAPT, are summarized in Table 1. Calculations are based on counterpoise-corrected RI-MP2/aVDZ optimized geometries. In order to estimate the errors in the above calculations, we have performed counterpoise-corrected optimization at the RI-MP2/cc-pVTZ and RI-MP2/aug-cc-pVTZ levels. RI-MP2/CBS and CCSD(T)/CBS energies are also estimated by using aVTZ–aVQZ extrapolation (Table 1, Table S4 of the Supporting Information). We find that the results from the counterpoise-corrected optimization at the RI-MP2/cc-pVTZ level for the most stable dimers (Py-D, Pz-D, Tr-S, and Tt- T^N) are closer to those at the RI-MP2/aug-cc-pVDZ optimization. R_c and R_v have an average absolute error of 0.01 Å, while R_d and angle parameters show no significant changes. For the most stable displaced-stacked and T-shaped dimers at the counterpoise-corrected optimization at the RI-MP2/aug-cc-pVTZ level, R_c and R_v have an average absolute error of 0.07 Å, while R_d and angle parameters show no significant changes. RI-MP2/CBS (most stable dimers) and CCSD(T)/CBS energies (Tr-D and Tt-

Table 1. RI-MP2/CBS ($-E_{MP2}$) and CCSD(T)/CBS ($-E_{tot}$) Binding Energies Using the aVDZ–aVTZ Extrapolation and SAPT Energy Components (in kcal/mol) of the Selected Low-Energy Dimers (with the Lowest Energy in Bold for Each $\#_N$) at the BSSE-Corrected RI-MP2/aug-cc-pVDZ Geometries^a

	$-E_{MP2}$	$-E_{tot}$	$-E_{es}$	$-E_{in}$	$-E_{dp}$	E_x
Stacked						
Bz-S	3.38	1.53	0.71	0.32	5.96	5.46
Py-S	4.64	2.79	2.21	0.37	6.38	6.17
Pz-S	5.74	3.48	3.42	0.31	6.93	7.18
Tr-S	5.00 [5.13]	3.92	2.75	0.40	6.94	6.18
Tt-S	5.64	3.13	2.31	0.44	7.14	6.76
Displaced-Stacked						
Bz-D	4.93	2.62	2.92	0.96	7.89	9.14
Py-D	6.19 [6.20]	3.80	4.48	0.99	8.11	9.78
Pz-D	6.76 [6.88]	4.09	4.94	1.06	8.48	10.39
Tr-D	5.22 [5.34]	3.82 [3.73]	3.00	0.52	7.39	7.09
Tt-D	6.47 [6.62]	3.67	3.46	1.19	8.25	9.23
T-Shaped						
Bz-T ^H	3.72	2.84	2.15	0.64	4.63	4.57
Py-T ^H	4.46 [4.52]	3.56 ^b	3.23	0.78	4.83	5.29
Pz-T ^H	3.50	2.70	2.20	0.65	4.42	4.56
Tr-T ^H	2.65	2.22	1.51	0.63	3.94	3.87
Tt-T ^H	1.40	0.91	-0.60	0.42	3.03	1.93
Py-T ^N	3.53	2.57	2.62	1.01	4.84	5.89
Pz-T ^N	4.65 [4.72]	3.36	4.76	0.93	5.23	7.57
Tr-T ^N	4.18 [4.25]	3.44	4.44	0.68	4.77	6.45
Tt-T ^N	5.70 [5.80]	4.27 [4.25]	6.19	1.11	4.66	7.69

^a Values in brackets are calculated using the aVTZ–aVQZ extrapolation calculated at the BSSE corrected RI-MP2/aug-cc-pVTZ geometries. ^b Py-T^H is nearly isoenergetic to Py-D.

T^N) derived from the aVDZ–aVTZ extrapolation have average absolute errors of 0.09 and 0.04 kcal/mol, respectively, in comparison with that derived from the aVTZ–aVQZ extrapolation. Thus, there is no significant difference in geometrical parameters and energies with increasing size of basis sets, and hence subsequent discussion will be based on the counterpoise-corrected RI-MP2/aVDZ optimized geometries and CBS energies obtained from the aVDZ–aVTZ extrapolation.

Variation of the binding energy and individual energy components with increasing $\#_N$ is shown in Figure 3. The dominant attractive contribution in the binding energy of benzene with either substituted monomers²¹ or heterocycles⁶ is dispersion, which increases with increasing substitution. The dispersion effects would render electrostatic contributions to the total binding energy (E_{tot}) less significant. The situation is different for the present system and further complicated due to the presence of multipole interactions. For the S series (Bz-S \rightarrow Py-S \rightarrow Pz-S \rightarrow Tr-S \rightarrow Tt-S, solid lines), the binding energy increases only from Bz-S to Tr-S. While the variation in induction binding energy ($-E_{in} = 0.7 \pm 0.4$ kcal/mol) is small for all cases, the large dispersion binding energy ($-E_{dp}$) levels off at Pz-S. On the other hand, the electrostatic binding energy ($-E_{es}$) becomes larger from Bz-S to Pz-S but decreases thereafter. A similar trend is noted for the exchange repulsion energy (E_x) except for a particularly small value for Tr-S. Individual energy contributions do not correlate well with E_{tot} except for E_{es} and E_{dp} , which have $R^2 = 0.84$ and 0.76, respectively (E_{in} , $R^2 = 0.17$; E_x , $R^2 = 0.48$).

Displaced-stacked structures of the D series (Bz-D \rightarrow Py-D \rightarrow Pz-D \rightarrow Tr-D \rightarrow Tt-D, dashed lines) are more stable than

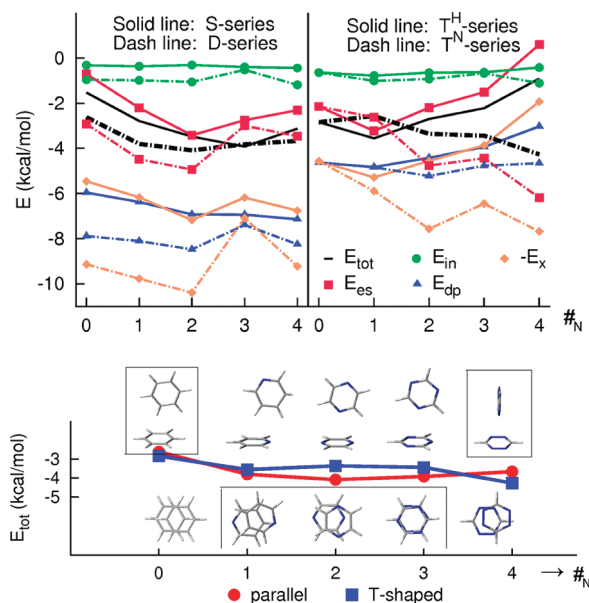


Figure 3. Interaction binding energies and energy components as a function of the number of N atoms ($\#_N$) in a series of dimer complexes. Lowest energy dimers are enclosed in boxes.

the stacked S-series structures, with the exception of Tr-D, which is almost isoenergetic to Tr-S. Stability increases from Bz-D to Pz-D but decreases from Pz-D to Tt-D. As compared with Tz-D, Tr-D has a smaller $-E_{es}$ and $-E_{dp}$ but is more stable due to a much smaller E_x . Correlation of each energy component with the total binding energy is poor ($R^2 = 0.44, 0.07, 0.01$, and 0.001 for E_{es} , E_{dp} , E_x , and E_{in} , respectively), reflecting the mixing of E_{es} , E_{dp} , and E_x .

A significant contribution to the stability of the T-shaped isomer of the Bz dimer is C–H $\cdots\pi$ interaction. Nitrogen enhances electrostatic interaction by withdrawing electron density from the ortho or para H, thus increasing its partial positive charge.⁷ However, this is countered by the resulting distortion in the π -electron cloud of the ring, which consequently weakens the C–H $\cdots\pi$ interaction. This trend was observed following the T^H series (Bz-T^H \rightarrow Py-T^H \rightarrow Pz-T^H \rightarrow Tr-T^H \rightarrow Tt-T^H; solid lines in Figure 3), with Tt-T^H having a repulsive electrostatic term. Electrostatic effects become more favorable along the T^N-series (Py-T^N \rightarrow Pz-T^N \rightarrow Tr-T^N \rightarrow Tt-T^N, dash lines), where the orientation is characterized by an N atom pointing toward the positively charged or electron-deficient ring center. In the T^N series, the increase in $-E_{es}$, however, is tempered by E_x as the vertical ring distance becomes shorter. The change in E_{dp} is relatively small from Py to Tt as the distance from the N atom to the ring center is similar. The binding energy for the T^N series correlates well with the electrostatic energy ($R^2 = 0.99$) and poorly with other contributing energy terms. The complexes become more stabilized/destabilized with increasing $\#_N$ for the T^N/T^H series. For the T^H series, all energy components have a good correlation with the total binding energy with $R^2 \rightarrow \sim 0.9$.

Predicted geometries and energetics reasonably explain the N-containing heterocyclic pairs in organic crystals. The small difference in binding energy of the displaced-stacked and T-shaped isomers for Py pairs is demonstrated by nearly equal distributions in the displaced-stacked and T-shaped regions in

Figure 1. Displaced-stacked Py and stacked Tr pairs have antiparallel dipoles and staggered quadrupole orientations. A Tt pair shows a T-shaped structure. Stable displaced-stacked Pz pairs are observed in the gas phase⁹ (Supporting Information).

Conclusion

In summary, heterocyclic dimers preferentially form a stacked/displaced-stacked arrangement, except for Tt, since Tt-T^N is more stable than Tt-D. Displaced-stacked isomers are more stable than the stacked ones except for Tr. For T-shaped isomers, the most stable Py-T^H has C-H $\cdots\pi$ interaction but changes to N $\cdots\pi$ interaction in the cases of Pz-T^N, Tr-T^N, and Tt-T^N. Dispersion effects dominate, particularly for stacked/displaced-stacked conformers. But, relative stabilities can be inferred mostly from the electrostatic contribution as envisaged by its better correlation with binding energies of the complexes except for the displaced-stacked conformers which are governed by E_{es} , E_{dp} , and E_{x} in a complicated manner.²¹ The present understanding would be very useful for designing diverse characteristic molecular models for intriguing molecular assembling and engineering.

Acknowledgment. This work was supported by NRF (WCU: R32-2008-000-10180-0; EPB Center: 2009-0063312, BK21, GRL) and KISTI (KSC-2008-K08-0002).

Supporting Information Available: Geometrical parameters and energies of all of the calculated dimer complexes along with CSD search data analysis and complete references. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) (a) Hunter, C. A.; Sanders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525–5534. (b) Hobza, P.; Selzle, H. L.; Schlag, E. W. *Chem. Rev.* **1994**, *94*, 1767–1785. (c) Kim, K. S.; Tarakeshwar, P.; Lee, J. Y. *Chem. Rev.* **2000**, *100*, 4145–4185. (e) Singh, N. J.; Min, S. K.; Kim, D. Y.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 515–529.
- (2) Singh, N. J.; Lee, H. M.; Hwang, I.-C.; Kim, K. S. *Supramol. Chem.* **2007**, *19*, 321–332.
- (3) Burley, S. K.; Petsko, G. A. *Science* **1985**, *229*, 23–28.
- (4) Cerny, J.; Kabelac, M.; Hobza, P. *J. Am. Chem. Soc.* **2008**, *130*, 16055–16059.
- (5) (a) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10888. (b) Tsuzuki, S.; Honda, K.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104–112. (c) Lee, E. C.; Hong, B. H.; Lee, J. Y.; Kim, J. C.; Kim, D.; Kim, Y.; Tarakeshwar, P.; Kim, K. S. *J. Am. Chem. Soc.* **2005**, *127*, 4530–4537. (d) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem.* **2006**, *110*, 10656–10668. (e) Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Phys. Chem. A* **2006**, *110*, 10345–10354. (f) Kim, E.; Paliwal, S.; Wilcox, C. S. *J. Am. Chem. Soc.* **1998**, *120*, 11192–11193. (g) Ren, R.; Jin, Y.; Kim, K. S.; Kim, D. H. *J. Biomol. Struct. Dyn.* **1997**, *15*, 401–405. (h) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100*, 18790–18794.
- (6) Wang, W.; Hobza, P. *Chem. Phys. Chem.* **2008**, *9*, 1003–1009.
- (7) Hohenstein, E. G.; Sherrill, C. D. *J. Phys. Chem. A* **2009**, *113*, 878–886.
- (8) (a) Piacenza, M.; Grimme, S. *Chem. Phys. Chem.* **2005**, *6*, 1554–1558. (b) Mishra, B. K.; Sathyamurthy, N. *J. Phys. Chem. A* **2005**, *109*, 6–8.
- (9) Busker, M.; Svartsov, Y. N.; Haber, T.; Kleinermanns, K. *Chem. Phys. Lett.* **2009**, *467*, 255–259.
- (10) Mignon, P.; Loverix, S.; Geerlings, P. *Chem. Phys. Lett.* **2005**, *401*, 40–46.
- (11) Šponer, J.; Hobza, P. *Chem. Phys. Lett.* **1997**, *267*, 263–270.
- (12) Bates, D. M.; Anderson, J. A.; Oloyede, P.; Tschumper, G. S. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2775–2779.
- (13) (a) Hong, B. H.; Lee, J. Y.; Lee, C.-W.; Kim, K. C.; Bae, S. C.; Kim, K. S. *J. Am. Chem. Soc.* **2001**, *123*, 10748–10749. (b) Lee, J. Y.; Hong, B. H.; Kim, W. Y.; Min, S. K.; Kim, Y.; Jouravlev, M. V.; Bose, R.; Kim, K. S.; Hwang, I.-C.; Kaufman, L. J.; Wong, C. W.; Kim, P.; Kim, K. S. *Nature* **2009**, *460*, 498–501.
- (14) (a) Janiak, C. *J. Chem. Soc., Dalton Trans.* **2000**, 3885–3896. (b) Sony, S. M. M.; Ponnuswamy, M. N. *Cryst. Growth Des.* **2006**, *6*, 736–742.
- (15) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. *Chem. Phys. Lett.* **1993**, *208*, 359–363.
- (16) (a) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646. (b) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208–1221.
- (17) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.
- (18) Bukowski, R. et al. *SAPT2008*; University of Delaware: Newark, DE, 2008. See also ref 17.
- (19) *TURBOMOLE V6.02009*; University of Karlsruhe and Forschungszentrum Karlsruhe GmbH: Karlsruhe, Germany, 2007. Available from <http://www.turbomole.com> (accessed Jun 2010).
- (20) Werner, H.-J. et al. *MOLPRO*, version 2006.1; University College Cardiff Consultants Limited: Cardiff, Wales, U.K., 2006. See <http://www.molpro.net> (accessed Jun 2010).
- (21) (a) Ringer, A. L.; Sherrill, C. D. *J. Am. Chem. Soc.* **2009**, *131*, 4574–4575. (b) Lee, E. C.; Kim, D.; Jurecka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446–3457. (c) Wheeler, S. E.; Houk, K. N. *J. Am. Chem. Soc.* **2009**, *131*, 4574–4575.

CT100182U

Optimal Weights in Serial Generalized-Ensemble Simulations

Riccardo Chelli*

*Dipartimento di Chimica, Università di Firenze, Via della Lastruccia 3,
I-50019 Sesto Fiorentino, Italy and European Laboratory for Nonlinear Spectroscopy
(LENS), Via Nello Carrara 1, I-50019 Sesto Fiorentino, Italy*

Received February 22, 2010

Abstract: In serial generalized-ensemble simulations, the sampling of a collective coordinate of a system is enhanced through non-Boltzmann weighting schemes. A popular version of such methods is certainly the simulated tempering technique, which is based on a random walk in temperature ensembles to explore the phase space more thoroughly. The most critical aspect of serial generalized-ensemble methods with respect to their parallel counterparts, such as replica exchange, is the difficulty of weight determination. Here we propose an adaptive approach to update the weights on the fly during the simulation. The algorithm is based on generalized forms of the Bennett acceptance ratio and of the free energy perturbation. It does not require intensive communication between processors and, therefore, is prone to be used in distributed computing environments with modest computational cost. We illustrate the method in a series of molecular dynamics simulations of a model system and compare its performances to two recent approaches, one based on adaptive Bayesian-weighted histogram analysis and the other based on initial estimates of weight factors obtained by potential energy averages.

1. Introduction

In computer simulations of complex systems it is often difficult to obtain accurate canonical distributions by conventional Boltzmann sampling because simulated systems tend to get trapped in local minimum-energy states. A strategy to tackle the problem is to perform simulations using non-Boltzmann probability weight factors, so that a random walk in energy space can be realized. In this context, a new class of simulation algorithms, generically termed generalized-ensemble algorithms,¹ has been developed. In the multicanonical approach,^{2,3} for instance, phase space is sampled with a probability proportional to an approximate estimate of the inverse potential energy density of states. In the simulated tempering (ST) technique,^{4,5} weighted sampling is used to produce a random walk in temperature space thus allowing the system to overcome energy barriers. An important limitation of ST is that an evaluation of the free energy as a function of temperature is needed as input to ensure equal visitation of temperatures, and eventually a

faster convergence of structural properties.⁶ The temperature replica exchange method^{7–10} (REM), also known as parallel tempering, was developed as an evolution of ST to eliminate the need to know a priori temperature-dependent free energies. Many other methodologies and combinations thereof have also been proposed,^{1,11–19} including approaches based on nonrandom walks in the ensemble space.^{20,21}

The idea of ST and temperature-REM can be readily extended to other ensemble parameters (e.g., pressure, interatomic distances, torsional bond angles, switching coordinate in alchemical transformations, etc.). The term generalized-ensemble, used to refer to such methods, arises from this generalization. The further classification of *serial* generalized-ensemble (SGE) and *parallel* generalized-ensemble algorithms is also used to distinguish between schemes based on single-replica transitions (like in ST) and on synchronous double-replica transitions (like in REM), respectively.²² Among generalized-ensemble algorithms, ST and temperature-REM allow an extensive exploration of phase space without configurational restraints. This gives the possibility of recovering not only the global minimum-energy

* Author e-mail: riccardo.chelli@unifi.it.

state but also any equilibrium thermodynamic quantity as a function of temperature. The potential of mean force (PMF)^{23,24} along a chosen collective coordinate can also be computed a posteriori by multiple-histogram reweighting techniques.^{25,26} In this case, however, many configurations sampled at high temperatures will give small contribution to the PMF at low (ordinary) temperature with the result of making quite ineffective the algorithm. PMF calculation is instead improved by performing generalized-ensemble canonical simulations in the space of the collective coordinate (for example, the space of the end-to-end distance of a biopolymer). In such a case, all system configurations will contribute equally to construct the PMF at the given temperature.¹⁷

Comparisons between ST and temperature-REM have been reported recently.^{6,27,28} The overall conclusions of these studies are that ST consistently gives a higher rate of delivering the system between high- and low-temperature states as well as a higher rate of transversing the potential energy space. Moreover, ST is well-suited to distributed computing environments because synchronization and communication between replicas/processors can be avoided. On the other side, an effective application of ST and, in general, of SGE methods requires a uniform exploration of the ensemble space. In order to satisfy this criterion, acceptance rates must be not only high but also symmetric between forward and backward directions of the ensemble space. This symmetry can be achieved by performing weighted sampling, where weights are correlated with the dimensionless free energies of the ensembles. The knowledge of such free energies is not needed in parallel generalized-ensemble methods because replica exchanges occur between microstates of the same extended thermodynamic ensemble. To achieve rapid sampling of the ensemble space through high acceptance rates, we need to choose ensembles appropriately so that neighboring ensembles overlap significantly. This last requirement is common to both SGE and parallel generalized-ensemble methods and in general does not depend on the specific algorithm used in simulation. Therefore the most critical aspect in applying SGE schemes is the determination of weight factors (viz. dimensionless free energy differences between neighboring ensembles). This issue has been the subject of many studies, especially addressed to ST simulations. The first attempts are based on short trial simulations.^{5,29,30} The proposed procedures are however quite complicated and computationally expensive for systems with many degrees of freedom. Later, Mitsutake and Okamoto suggested to perform a short REM simulation to estimate ST weight factors³¹ via multiple-histogram reweighting.^{25,26} A further approximated, but very simple, approach to evaluate weight factors is based on average energies calculated by means of conventional molecular dynamics simulations.²² The weight factors obtained by the average-energy method²² were later demonstrated to correspond to the first term of a cumulant expansion of free energy differences.²⁷ Huang et al. used approximated estimates of potential energy distribution functions (from short trial molecular dynamics simulations) to equalize the acceptance rates of forward and backward transitions between

neighboring temperatures, ultimately leading to a uniform temperature sampling in ST.³² The techniques illustrated above have been devised to determine weight factors to be used without further refinement³¹ or as an initial guess to be updated during the simulation.^{22,32} In the former case, these approximate factors should (hopefully) guarantee an almost random walk through the ensemble space. However, as remarked in ref 6, the estimate of accurate weight factors may be very difficult for complex systems. Inaccurate estimates, though unaffected the basic principles of SGE methods, do affect the sampling performances in terms of simulation time needed to achieve convergence of structural properties.⁶

As discussed above, dimensionless free energy differences between ensembles (viz. weight factors) may also be the very aim of the simulation.¹⁷ In such cases, accurate determination of weight factors is not simply welcome but necessary. This can be done a posteriori using multiple-histogram reweighting techniques^{25,26} or using more or less efficient updating protocols applied during the simulation.^{6,19,32–34}

In this article we present an adaptive method to calculate weight factors in SGE simulations based on generalized expressions^{35,36} of the Bennett method³⁷ and of the free energy perturbation.³⁸ Although the method may appear as a downgrading of the multiple-histogram reweighting algorithm,^{25,26} it is asymptotically exact and requires a low computational time per updating step. Moreover, since the overlap between the distribution functions of the generalized dimensionless work³⁶ spent in the forward and backward transitions between neighboring ensembles must be not negligible, the accuracy of the method is comparable to the multiple-histogram reweighting approach. The algorithm is suited not only to calculate the free energy on the fly during the simulation but also as a possible criterion to establish whether equilibration has been reached. We illustrate the method on a model system made of two particles interacting through a double-well potential and solvated by a monatomic fluid. This model system contains much of condensed-phase physics and may be viewed as an elementary example of molecular docking with an energy barrier between the initial and final states. SGE simulations in temperature space (ST simulations) and in the space of the interparticle distance are carried out. The performances of our algorithm in recovering free energies as a function of temperature and interparticle distance (i.e., the PMF) are compared with those of various approaches, including multiple-histogram reweighting as reformulated in ref 39, the recent Bayesian weighted histogram analysis method³⁴ (ABWHAM), and the method based on the initial estimates of the weight factors obtained by averaging the potential energy of short trial simulations.²²

The outline of the article follows. In Section 2, SGE methods are introduced. The algorithm for computing optimal weights is proposed in Section 3. Technical details on the simulations and on the system are given in Section 4, while the simulation results are reported and discussed in Section 5. Concluding remarks can be found in Section 6.

2. Introduction to Serial Generalized-Ensemble Methods

A SGE method deals with a set of N ensembles associated with different dimensionless Hamiltonians $h_n(x, p)$, where x and p denote the atomic coordinates and momenta of a microstate⁴⁰ and $n = 1, 2, \dots, N$ denotes the ensemble. Each ensemble is characterized by a partition function expressed as

$$Z_n = \int e^{-h_n(x,p)} dx dp \quad (1)$$

In ST simulations the dimensionless Hamiltonian is

$$h_n(x, p) = \beta_n H(x, p) \quad (2)$$

where $H(x, p)$ is the original Hamiltonian and $\beta_n = (k_B T_n)^{-1}$, with k_B being the Boltzmann constant and T_n the temperature of the n th ensemble. If we express the Hamiltonian as a function of λ , namely a parameter correlated with an arbitrary collective coordinate of the system (or even corresponding to the pressure), then the dimensionless Hamiltonian associated with the n th λ -ensemble is

$$h_n(x, p) = \beta H(x, p; \lambda_n) \quad (3)$$

Here all ensembles have the same temperature. It is also possible to construct a generalized ensemble for multiple parameters⁴¹ as

$$h_n(x, p) = \beta_n H(x, p; \lambda_i) \quad (4)$$

In this example two parameters, T and λ , are employed, but no restraint is actually given to the number of ensemble spaces. Generalized-ensemble algorithms have a different implementation dependent on whether the temperature is included in the collection of sampling spaces (eqs 2 and 4). Here we adhere to the most general context without specifying any form of $h_n(x, p)$, except when we discuss implementation of ST (Section 2.1) and of the PMF calculation (Section 2.2).

In SGE simulations, the probability of a microstate (x, p) in the n th ensemble [from now on denoted as $(x, p)_n$] is proportional to $\exp[-h_n(x, p) + g_n]$, where g_n is a factor, different for each ensemble, that must ensure almost equal visitation of the N ensembles. The extended partition function of this “system of ensembles” is

$$Z = \sum_{n=1}^N \int e^{-h_n(x,p)+g_n} dx dp = \sum_{n=1}^N Z_n e^{g_n} \quad (5)$$

where Z_n is the partition function of the system in the n th ensemble (eq 1). In practice SGE simulations work as follows. A single simulation is performed in a specific ensemble, say n , using Monte Carlo or molecular dynamics sampling protocols, and after a certain interval, an attempt is made to change the microstate $(x, p)_n$ to another microstate of a different ensemble $(x', p')_m$. Since high acceptance rates are obtained as the ensembles n and m overlap significantly, the final ensemble m is typically close to the initial one, namely $m = n \pm 1$.⁴² In principle, the initial and final microstates can be defined by different coordinates and/or

momenta ($x \neq x'$ and/or $p \neq p'$), though the condition $x = x'$ is usually adopted. The transition probabilities for moving from $(x, p)_n$ to $(x', p')_m$ and vice versa have to satisfy the detailed balance condition:

$$P_n(x, p)P(n \rightarrow m) = P_m(x', p')P(m \rightarrow n) \quad (6)$$

where $P_n(x, p)$ is the probability of the microstate $(x, p)_n$ in the extended canonical ensemble (eq 5):

$$P_n(x, p) = Z^{-1} e^{-h_n(x,p)+g_n} \quad (7)$$

In eq 6, $P(n \rightarrow m)$ is a shorthand for the conditional probability of the transition $(x, p)_n \rightarrow (x', p')_m$, given the system is in the microstate $(x, p)_n$ [with analogous meaning of $P(m \rightarrow n)$]. Using eq 7 together with the analogous expression for $P_m(x', p')$ in the detailed balance and applying the Metropolis's criterion, we find that the transition $(x, p)_n \rightarrow (x', p')_m$ is accepted with probability:

$$\text{acc}[n \rightarrow m] = \min(1, e^{h_n(x,p)-h_m(x',p')+g_m-g_n}) \quad (8)$$

The probability of sampling a given ensemble is

$$P_n = \int P_n(x, p) dx dp = Z_n Z^{-1} e^{g_n} \quad (9)$$

Uniform sampling sets the condition $P_n = N^{-1}$ for each ensemble ($n = 1, \dots, N$) that leads to the equality:

$$g_n = -\ln Z_n + \ln\left(\frac{Z}{N}\right) \quad (10)$$

Equation 10 implies that, to get uniform sampling, the difference $g_m - g_n$ in eq 8 must be replaced with $f_m - f_n$, where f_n is the dimensionless free energy related to the actual free energy of the ensemble n by the relation $f_n = \beta F_n = -\ln Z_n$, where β is the inverse temperature of the ensemble. Here we are interested in determining such free energy differences that will be referred as optimal weight factors, or simply, optimal weights. Accordingly, in the acceptance ratio we will use f_n instead of g_n .

2.1. SGE Simulations in Temperature-Space (Simulated Tempering). In SGE Monte Carlo simulations conducted in temperature space (ST simulations), eq 2 holds. Specifically, since only configurational sampling is performed, we have

$$h_n(x) = \beta_n V(x) \quad (11)$$

where $V(x)$ is the (potential) energy of the configuration x . Therefore, transitions from n to m ensemble, realized at fixed configuration, are accepted with probability:

$$\text{acc}[n \rightarrow m] = \min(1, e^{(\beta_n - \beta_m)V(x) + f_m - f_n}) \quad (12)$$

When the system evolution is performed with molecular dynamics simulations, the situation is slightly more complicated. Suppose we deal with canonical ensembles (to simplify the treatment and the notation we consider constant-volume and constant-temperature ensembles, though extension to constant-pressure and constant-temperature ensembles is straightforward). Usually, constant temperature is implemented through the Nosé–Hoover method^{43,44} or extensions

of it.⁴⁵ With the symbol p_i , we will denote the momentum conjugated to the dynamical variable associated with the thermostat. Also in this case eq 2 holds, but it takes the form

$$h_n(x, p, p_i) = \beta_n H(x, p, p_i) \quad (13)$$

In this equation, $H(x, p, p_i) = V(x) + K(p) + K(p_i)$ is the extended Hamiltonian of the system, where $V(x)$ is the potential energy, while $K(p)$ and $K(p_i)$ are the kinetic energies of the particles and thermostat, respectively. As in the Monte Carlo version, transitions from n to m ensemble are realized at fixed configuration, while particle momenta are rescaled as

$$\begin{aligned} p' &= p(T_m/T_n)^{1/2} \\ p'_i &= p_i(T_m/T_n)^{1/2} \end{aligned} \quad (14)$$

As in REM,⁸ the scaling drops the momenta out of the detailed balance, and the acceptance ratio takes the form of eq 12. Note that, if more thermostats are adopted,⁴⁵ then all additional momenta must be rescaled according to eq 14.

2.2. SGE Simulations in λ -Space. In SGE simulations conducted in a generic λ -space at constant temperature, the dimensionless Hamiltonian is given by eq 3. In our molecular dynamics simulations we use a Hamiltonian aimed to sample the distance between two target particles. There are several ways to model such a Hamiltonian. Our choice is

$$h_n(x, p, p_i) = \beta[H(x, p, p_i) + k(r - \lambda_n)^2] \quad (15)$$

where, as usual, $H(x, p, p_i)$ is the extended Hamiltonian. In eq 15, r is the instantaneous distance between the target particles, and k is a constant. As in ST simulations, transitions from n to m ensemble occur at fixed configuration. However, in this case, there is no need of rescaling momenta because they drop out of the detailed balance condition naturally. The resulting acceptance ratio is

$$\text{acc}[n \rightarrow m] = \min(1, e^{\beta k[(r - \lambda_n)^2 - (r - \lambda_m)^2] + f_m - f_n}) \quad (16)$$

The same ratio is obtained using Monte Carlo sampling. In this kind of simulation, the free energy as a function of λ corresponds to the biased PMF^{23,24} along the coordinate associated with λ . Biasing arises from the harmonic potential being added to the original Hamiltonian (see eq 15). However, reweighting schemes are available to recover the unbiased PMF along the real coordinate.^{25,26,46,47}

3. The Algorithm for Optimal Weights

3.1. Tackling Free Energy Estimates. The algorithm proposed to calculate the optimal weight factors, namely the dimensionless free energy differences between ensembles (see Section 2), is based on the Bennett acceptance ratio^{37,48} and on the free energy perturbation formula.³⁸ We start by showing that the difference between the dimensionless Hamiltonians appearing in the acceptance ratio (see eq 8) can be viewed as the generalized dimensionless work done on the system during the transition $(x, p)_n \rightarrow (x', p')_m$. The concept of generalized dimensionless work in systems subject to mechanical and thermal nonequilibrium changes has been

extensively discussed recently.^{35,36,49} In particular it has been shown (see eq 45 in ref 36) that, in a nonequilibrium realization performed with extended-Lagrangian molecular dynamics,⁵⁰ the generalized dimensionless work is

$$W = \beta_\tau H'(\tau) - \beta_0 H'(0) \quad (17)$$

where τ is the duration of the realization and

$$H'(\tau) = H(x, p, p_i) + k_B T_\tau \nu(x_i) \quad (18)$$

where $H(x, p, p_i)$ is defined in eq 13 and $\nu(x_i)$ is a linear function of the configurational variables x_i associated with the thermostat (see eq 42 in ref 36). For simplicity, in eq 18 we have only reported the explicit time dependence of the temperature. Moreover, we have considered to deal with thermal changes alone using constant-volume and constant-temperature equations of motion. Extending the treatment to constant-pressure and constant-temperature algorithms and to systems subject to generic λ , e.g. mechanical, changes is straightforward.³⁶ Note that, when no changes are externally applied to the system, H' is exactly the quantity conserved during the constant-volume and constant-temperature simulation. Accordingly, the work W is zero. The above definition of generalized dimensionless work is valid for arbitrary values of τ . In the special case of instantaneous thermal changes and variations of the microstate variables, as it occurs in ST simulations, the times 0 and τ in eq 17 refer to the states instantaneously before and after the $(x, p)_n \rightarrow (x', p')_m$ transition, respectively. Therefore, according to the notation introduced above, eq 17 can be rewritten as

$$W[n \rightarrow m] = \beta_m H(x', p', p'_i) - \beta_n H(x, p, p_i) + \nu(x'_i) - \nu(x_i) \quad (19)$$

where x_i and x'_i are the values of the configurational thermostat-variables before and after the $(x, p)_n \rightarrow (x', p')_m$ transition, respectively. In the first two terms on the right-hand side of eq 19, we can recognize the dimensionless Hamiltonians $h_m(x', p', p'_i)$ and $h_n(x, p, p_i)$. It is important to observe that, in generalized-ensemble simulations, an arbitrary change of x_i during a transition does not affect the acceptance ratio or the dynamics of the system. Therefore, by setting $x'_i = x_i$ and generalizing to λ changes, we recover the equality:

$$W[n \rightarrow m] = h_m(x', p', p'_i) - h_n(x, p, p_i) \quad (20)$$

This result is general and can be proved to be valid also for Monte Carlo simulations. Using $W[n \rightarrow m]$, the acceptance ratio of eq 8 becomes

$$\text{acc}[n \rightarrow m] = \min(1, e^{\Delta f_{n \rightarrow m} - W[n \rightarrow m]}) \quad (21)$$

where $\Delta f_{n \rightarrow m} = f_m - f_n$. The quantity $W[n \rightarrow m] - \Delta f_{n \rightarrow m}$ can be interpreted as the generalized dimensionless work dissipated in the transformation (see eq 17 in ref 36).

Until now we have simply restated the acceptance ratio of SGE simulations in terms of the generalized dimensionless work $W[n \rightarrow m]$. The truly important aspect of this treatment is that the knowledge of $W[n \rightarrow m]$ and $W[m \rightarrow n]$ stored during the sampling gives us the possibility of evaluating

the optimal weights $\Delta f_{n \rightarrow m}$ using the Bennett method³⁷ reformulated with maximum likelihood arguments.^{36,48} For example, in ST simulations we must take memory of the quantities $W[n \rightarrow m] = (\beta_m - \beta_n)V_n(x)$ and $W[m \rightarrow n] = (\beta_n - \beta_m)V_m(x)$, where the subscripts of the potential energy indicate the ensemble at which sampling occurs. Thus, for each pair of neighboring ensembles n and m , we generate two collections of “instantaneous generalized dimensionless works”: $W_1[m \rightarrow n], W_2[m \rightarrow n], \dots$, etc. and $W_1[n \rightarrow m], W_2[n \rightarrow m], \dots$, etc. Let us denote the number of elements of such collections with $N_{m \rightarrow n}$ and $N_{n \rightarrow m}$. So $\Delta f_{n \rightarrow m}$ can be calculated by solving the equation (see eq 27 in ref 36):

$$\sum_{i=1}^{N_{n \rightarrow m}} \left[1 + \frac{N_{n \rightarrow m}}{N_{m \rightarrow n}} e^{W_i[n \rightarrow m] - \Delta f_{n \rightarrow m}} \right]^{-1} - \sum_{j=1}^{N_{m \rightarrow n}} \left[1 + \frac{N_{m \rightarrow n}}{N_{n \rightarrow m}} e^{W_j[m \rightarrow n] + \Delta f_{n \rightarrow m}} \right]^{-1} = 0 \quad (22)$$

that just corresponds to the Bennett acceptance ratio for dimensionless quantities. It is important to point out that eq 22 is valid for nonequilibrium transformations, does not matter how far from equilibrium, and is rigorous only if the initial microstates of the transformations are drawn from equilibrium. Therefore care should be taken in verifying whether convergence/equilibrium is reached in the adaptive procedure. It should be noted that eq 22 is a straightforward generalization (to systems subject to thermal changes) of eq 8 in ref 48 that was specifically derived for systems subject to mechanical changes.

Shirts et al.⁴⁸ proposed a way of evaluating the square uncertainty (variance) of $\Delta f_{n \rightarrow m}$ from maximum likelihood methods by also correcting the estimate in the case of the restriction from fixed probability of forward and backward work measurements to fixed number of forward and backward work measurements. They provided a formula for systems subject only to mechanical work. However, by following the arguments in ref 36, it is straightforward to generalize the variance to a situation in which also thermal work is performed

$$\sigma^2(\Delta f_{n \rightarrow m}) = 2 \left\{ \sum_{i=1}^{N_{n \rightarrow m}} [1 + \cosh(W_i[n \rightarrow m] - \Delta f')]^{-1} + \sum_{j=1}^{N_{m \rightarrow n}} [1 + \cosh(W_j[m \rightarrow n] + \Delta f')]^{-1} \right\}^{-1} - N_{n \rightarrow m}^{-1} - N_{m \rightarrow n}^{-1} \quad (23)$$

where $\Delta f' = \Delta f_{n \rightarrow m} + \ln(N_{m \rightarrow n}/N_{n \rightarrow m})$. The quantity $\sigma^2(\Delta f_{n \rightarrow m})$ can be calculated once $\Delta f_{n \rightarrow m}$ is recovered from eq 22.

It is obvious that, in order to employ eq 22, both n and m ensembles must be visited at least one time. If statistics are instead retrieved from one ensemble alone, say n , then we have to resort to a different approach. The one we propose is consistent with the previous treatment. In fact, in the limit that only one work collection (specifically, the $n \rightarrow m$ collection) is available, eq 22 becomes⁴⁸ (compare with eq 21 in ref 36)

$$e^{-\Delta f_{n \rightarrow m}} = N_{n \rightarrow m}^{-1} \sum_{i=1}^{N_{n \rightarrow m}} e^{-W_i[n \rightarrow m]} \quad (24)$$

thus recovering the well-known fact that the free energy is the expectation value of the work exponential average.⁵¹

3.2. Implementation of Adaptive Free Energy Estimates in SGE Simulations. We now describe how the machinery introduced in Section 3.1 can be employed in the context of adaptive algorithms for SGE simulations. Suppose we deal with N ensembles of a generic Λ -space, be it a temperature space, a λ -space, or even a multiple-parameter space. Without loss of generality, we order the ensembles as $\Lambda_1 < \Lambda_2 < \dots < \Lambda_N$. Thus, $N - 1$ optimal weights, $\Delta f_{1 \rightarrow 2}, \Delta f_{2 \rightarrow 3}, \dots, \Delta f_{N-1 \rightarrow N}$, have to be estimated adaptively.

(1) At the beginning of the simulation we assign the system, i.e., the replica, to a randomly chosen ensemble and start the phase space sampling with the established simulation protocol (Monte Carlo or molecular dynamics). Note that several simulations may run in the generalized-ensemble space, each yielding an independent trajectory. Analogously to REM, a single simulated system will be termed “replica”. For the sake of simplicity, in the following presentation of the method we will take into account one replica alone. A discussion regarding multiple-replica simulations is reported in the final part of this section.

(2) Every L_a steps and for each ensemble n , we store into memory the quantities $W[n \rightarrow n + 1]$ and $W[n \rightarrow n - 1]$, computed as described in Section 3.1. There is no well-established recipe in choosing L_a , apart from the requirement that it should ensure (as large as possible) uncorrelation between work values. During the simulation we must also record the number of stored W elements, $N_{n \rightarrow n+1}$ and $N_{n \rightarrow n-1}$.

(3) Every L_b steps, such that $L_b \gg L_a$ (three orders of magnitude at least), we try a free energy update on the basis of eqs 22 or 24. The scheme we propose for $\Delta f_{n \rightarrow n+1}$ follows:

(a) First of all we check if the conditions $N_{n \rightarrow n+1} > N'$ and $N_{n+1 \rightarrow n} > N'$ are met. In such a case, eq 22 is applied (setting $m = n + 1$) using the stored dimensionless works (see point 2). The threshold N' is used as a control parameter for the accuracy of the calculation. Once $\Delta f_{n \rightarrow n+1}$ is known, its square uncertainty is computed according to eq 23. Then we set $N_{n \rightarrow n+1} = 0$ and $N_{n+1 \rightarrow n} = 0$ and cancel $W[n \rightarrow n + 1]$ and $W[n + 1 \rightarrow n]$ from computer memory. Whenever the free energy estimate and the correlated uncertainty are computed, the optimal weight to be used in the acceptance ratio (eq 21) is determined, applying standard formulas from maximum likelihood considerations (see Section 3.3). This step is realized for $n = 1, 2, \dots, N - 1$.

(b) If the criteria needed to apply eq 22 are not met and no $\Delta f_{n \rightarrow n+1}$ estimate is still available from point 3a, then we try to apply eq 24. In particular, two independent estimates of $\Delta f_{n \rightarrow n+1}$ are attempted. One comes from eq 24 by setting $m = n + 1$, whereas the other comes from eq 24 applied in the reverse direction (replace n with $n + 1$ and m with n in eq 24). The two estimates will be invoked in the acceptance ratio of $n \rightarrow n + 1$ and $n + 1 \rightarrow n$ ensemble transitions,

respectively (see next point 4). In the former case, we need to resort to additional arrays (denoted as $N_{n \rightarrow n+1}^{\text{up}}$ and $W^{\text{up}}[n \rightarrow n+1]$) to store $N_{n \rightarrow n+1}$ and $W[n \rightarrow n+1]$. Separate arrays are necessary because they are subject to different manipulation during the simulation. Specifically, if the condition $N_{n \rightarrow n+1}^{\text{up}} > N'$ is satisfied, then we calculate $\Delta f_{n \rightarrow n+1}$ via eq 24. This estimate is employed as such in the acceptance ratio. Then we set $N_{n \rightarrow n+1}^{\text{up}} = 0$ and cancel $W^{\text{up}}[n \rightarrow n+1]$ from computer memory. The same protocol is used to calculate $\Delta f_{n+1 \rightarrow n}$ from the quantities $N_{n+1 \rightarrow n}^{\text{down}}$ and $W^{\text{down}}[n+1 \rightarrow n]$. The additional arrays introduced here are updated as described in point 2. Note that in this procedure the arrays of step 3a are neither used nor changed. Note also that the procedure described here corresponds to the way of calculating the finite free energy differences in the free energy perturbation method.³⁸

- (c) If none of the above criteria is met, then optimal weights are not updated and conventional sampling continues. Storage of dimensionless works, as described at point 2, continues as well.

We point out that, if equilibrium is reached slowly (as in the case of large viscous systems or systems with very complex free energy landscape), then the replicas may tend to get trapped in limited regions of the ensemble space at the early stages of the simulation. This is basically due to initially inaccurate determination of $\Delta f_{n \rightarrow n+1}$ from eq 22 (point 3a). If such an event occurs, then subsequent free energy estimates from eq 22 may become very rare or even impossible. However, we can prevent this unwanted situation by passing to the updating criteria of point 3b when the criteria of point 3a are not met for a given (prior established) number of consecutive times. When equilibrium will be approached, the criteria of point 3b will favor transitions of the replicas between neighboring ensembles (this issue will be discussed in Section 5.3) and eventually the conditions to apply again the criteria of point 3a.

(4) Every L_c steps, a transition $(x, p)_n \rightarrow (x, p')_{n \pm 1}$ is attempted on the basis of the acceptance ratio of eq 21 and of the current value of $\Delta f_{n \rightarrow n \pm 1}$ (properly reweighted according to the equations reported in Section 3.3). If the estimate of $\Delta f_{n \rightarrow n \pm 1}$ is still not available from the methods described at points 3a and 3b, then the transition is not realized. The upward and downward transitions are chosen with equal probability. If the transition is accepted and the sampling occurs in the temperature space using molecular dynamics, then the momenta/velocities of the extended system are rescaled according to eq 14.

It is worthwhile stressing again that the procedures of point 3b are only aimed to furnish a reliable evaluation of optimal weights when such factors are still not available from the bidirectional algorithm (point 3a) or when the system is trapped in one or few ensembles (point 3c). Moreover, we remark that the free energy differences estimated via eq 24 tend to give larger acceptance rates in comparison to the exact free energy differences, thus favoring the transitions

toward the ensemble that has not been visited. This is a well-known (biasing) effect of exponential averaging,⁵² leading to a mean dissipated (dimensionless) work artificially low. As a matter of fact, this is a positive effect since it makes easier ensemble transitions during the equilibration phase of the simulation. This aspect will be further discussed in Section 5.3.

In the above discussion, we have not mentioned the number M of (independent) replicas that may run in the space of the N ensembles. In principle, M can vary from 1 to ∞ on the basis of our computer facilities. The best performance is obtainable if a one-to-one correspondence exists between replicas and computing processors. A rough parallelization could be obtained performing M independent simulations and then drawing the data from replicas at the end of the simulation to get augmented statistics. However, the calculation of the optimal weights would be much improved if they were periodically updated on the fly on the basis of the data drawn from all replicas. This is just what we do. In this respect, we notice that our version of multiple-replica SGE algorithm is prone to work efficiently also in distributed computing environments. The phase of the simulation where information is exchanged is that described at point 3 (free energy calculation). It should be noted that, when a free energy estimate is performed, the work arrays stored for each replica/processor (see point 2) do not need to be communicated to all other replicas/processors. Only the sums $\sum_{i=1}^{N_{n \rightarrow m}} [\cdot]^{-1} - \sum_{j=1}^{N_{m \rightarrow n}} [\cdot]^{-1}$ (case of eq 22), $\sum_{i=1}^{N_{n \rightarrow m}} [\cdot]^{-1} + \sum_{j=1}^{N_{m \rightarrow n}} [\cdot]^{-1}$ (case of eq 23), and $\sum_{i=1}^{N_{n \rightarrow m}} \exp(-W_i[n \rightarrow m])$ (case of eq 24), together with $N_{n \rightarrow m}$ and $N_{m \rightarrow n}$, must be exchanged for all $N - 1$ ensemble transitions. Then each replica/processor “will think by itself” to reassemble the global sums. Exchanging one information implies to send $M(M - 1)(N - 1)$ real/integer numbers through the net (~ 60 kB of information using 20 replicas and slightly less than 1 MB of information using 50 replicas). Only in the case of the iterative procedure of eq 22, one information has to be sent several times per free energy calculation (i.e., the number of iterations needed for solving the equation). The computational cost arising from computer communications can however be reduced updating the free energy rarely. Furthermore, in order to improve the first free energy estimate and hence to speed up the convergence, the M simulations should be started by distributing the replicas among neighboring ensembles, namely replica 1 to Λ_1 , replica 2 to Λ_2 , and so on. In the remainder of this paper, we will refer to the algorithm described in this section as BAR-SGE.

3.3. Free Energy Evaluation from Independent Estimates and Associated Variances. As discussed in Section 3.2, during a SGE simulation, optimal weights are evaluated using eq 22, and only temporary values are obtained from eq 24. Therefore, for each optimal weight, the simulation produces a series of estimates, $\Delta f_1, \Delta f_2, \dots, \Delta f_P$. At a given time, the current value of P depends, on average, on the time and the update frequency of optimal weights. In this section, for convenience, the subscript in Δf_i labels independent estimates. We also know that each Δf_i value is affected by an uncertainty quantified by the associated variance $\delta^2(\Delta f_i)$ calculated via eq 23. We can then write $\Delta \hat{f}$, the optimal

estimator of $P^{-1}\sum_{i=1}^P\Delta f_i$, by a weighted sum of the individual estimates:⁵³

$$\hat{\Delta f} = \frac{\sum_{i=1}^P [\delta^2(\Delta f_i)]^{-1} \Delta f_i}{\sum_{j=1}^P [\delta^2(\Delta f_j)]^{-1}} \quad (25)$$

Note that independent estimates with smaller variances have greater weight, and if the variances are equal, then the estimator $\hat{\Delta f}$ is simply the mean value of the estimates. The uncertainty in the resulting estimate can be computed from the variances of the single estimates as

$$\delta^2(\hat{\Delta f}) = \left\{ \sum_{j=1}^P [\delta^2(\Delta f_j)]^{-1} \right\}^{-1} \quad (26)$$

4. Details on Methods and System

We illustrate the BAR-SGE method on two series of simulations, one performed in the temperature space (ST simulations) and the other in the space of the distance between two particles, denoted as λ -space. In both cases, the calculations have been carried out on a model system made of two ‘‘solute’’ particles immersed into a Lennard-Jones fluid of 1398 (‘‘solvent’’) particles. Additional ST simulations have been performed on a larger sample made of two solute particles and 13 998 solvent particles. The solute particles interact each other through a double-well potential whose expression is

$$V(x) = 6[(x - 1)^2 - 0.1](x - 3)^2 \quad (27)$$

where $x = |x_2 - x_1|$ is the X component of the interparticle distance vector. Here and in the following all quantities are in reduced units. The solute particles are also constrained to move along the X direction through a combination of stiff harmonic potentials: $k_{yz}(y_1^2 + z_1^2 + y_2^2 + z_2^2)$, where (x_1, y_1, z_1) and (x_2, y_2, z_2) are the Cartesian coordinates of the particles and $k_{yz} = 5 \times 10^3$. With such a stiff potential, the quantity x appearing into eq 27 well approximates the actual interparticle distance, eventually eliminating the Jacobian contribution from the PMF along the interparticle direction. The same mass is used for both solute and solvent particles. Unitary Lennard-Jones parameters are employed for solute–solvent and solvent–solvent interactions, while only $V(x)$ accounts for the solute–solute interaction. All simulations have been carried out in constant-volume and constant-temperature ensembles using a cubic box with standard periodic boundary conditions. The density is 0.85, while the temperature is kept fixed by means of the Nosé–Hoover chain technique⁴⁵ with four coupled thermostats. Lennard-Jones interactions are cut off smoothly in the 3.0–3.5 distance range by multiplying the potential energy by a function $s(r)$ such that $s(r) = 1$ for $r \leq 3$, $s(r) = 0$ for $r \geq 3.5$, and $s(r) = 16r^3 - 156r^2 + 504r - 539$ for $3 < r < 3.5$. The time step (t -step) used in the small-sample simulations is $\sim 9.15 \times 10^{-3}$, while in the large-sample simulations t -step is $\sim 1.373 \times 10^{-2}$. For a given replica, initial positions of the solvent particles are random, while the solute particles

are taken with coordinates (0, 0, 0) and (0.5, 0, 0) in ST simulations and (0, 0, 0) and $(\lambda_n, 0, 0)$ in λ -space SGE simulations, where λ_n is the specific λ value associated with the ensemble from which the replica starts the dynamics.

Small-sample ST simulations have been carried out using 15 ensembles covering the temperature interval 0.6–1.2. The temperatures are spaced out on the basis of uniform steps of T^{-1} , namely $T_n^{-1} - T_{n+1}^{-1} = 5.95 \times 10^{-2}$. In large-sample simulations the same interval of temperature has been taken. However preliminary simulations have revealed that the above distribution of temperature provides negligible acceptance ratios. In order to get acceptance ratios greater than 10%, 30 ensembles/temperatures have been found necessary. Moreover it has been shown⁵⁴ that a better efficiency in terms of acceptance ratios is obtainable by distributing the temperature on the basis of the rule $T_{n+1} = aT_n$, where a is a constant dependent on the number of ensembles/temperatures and on the difference between maximum and minimum temperatures (in our case $a = 1.02419$). The acceptance ratio for ST simulations is given by eq 12.

SGE simulations in the λ -space have been carried out using 21 ensembles at $T = 0.6$ covering the distance interval 0.5–3.5 with a constant step size, $\lambda_{n+1} - \lambda_n = 0.15$. In this case, the acceptance ratio is given by eq 16 with a force constant k of 25. The k value has been chosen on the basis of short preliminary simulations to ensure overlap between neighboring ensembles.

All SGE and small-sample ST simulations have been carried out for a time of 1.5×10^6 t -steps per replica, while the large-sample ST simulations have been carried out for a time of 10^5 t -steps per replica. The various replicas in multiple-replica simulations are initially distributed in order of increasing temperature (ST simulations) or increasing λ (λ -space SGE simulations). Other details, such as the number of replicas M and the relevant parameters L_a, L_b, L_c , and N' (see Section 3.2), will be reported below.

5. Applications

5.1. Simulated Tempering Simulations. *5.1.1. Small-Sample Case.* In the context of ST, we report on the results of four multiple-replica simulations differing in the number of replicas, i.e., $M = 1, 5, 10$, and 15. The simulation parameters in t -step units are $L_a = 2, L_b = 2000, L_c = 10$, and $N' = 1000$ (see Section 3.2 for details). Note that, in the following, the 0 time corresponds to the starting random configuration, generated as described in Section 4. In Figure 1 we report four representative optimal weights, $\Delta f_{1-2}, \Delta f_{6-7}, \Delta f_{10-11}$, and Δf_{14-15} , as a function of time per replica (only the values computed by eq 22 are actually reported). These weights are associated with the temperature transitions $T_1 = 0.600 \rightleftharpoons T_2 = 0.622, T_6 = 0.730 \rightleftharpoons T_7 = 0.764, T_{10} = 0.884 \rightleftharpoons T_{11} = 0.933$, and $T_{14} = 1.120 \rightleftharpoons T_{15} = 1.200$. In Figure 1, the optimal weights calculated using the multiple Bennett acceptance ratio (MBAR) estimator³⁹ are also plotted. MBAR is equivalent to the multiple-histogram reweighting method^{25,26} in the limit that histogram bin widths are shrunk to 0 and corresponds to the Bennett acceptance ratio (eq 22) when only two states are considered. The

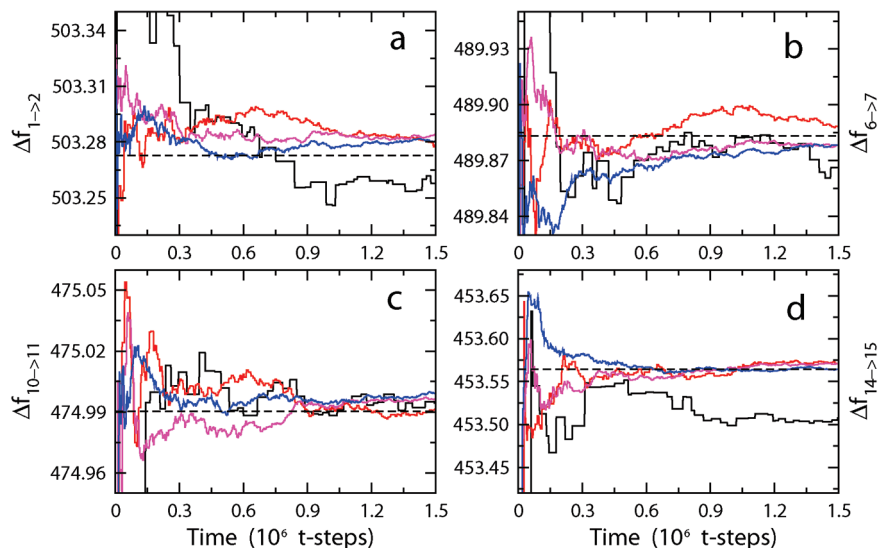


Figure 1. Representative BAR-SGE optimal weights as a function of time per replica obtained from small-sample ST simulations. Panels a–d: $\Delta f_{1 \rightarrow 2}$, $\Delta f_{6 \rightarrow 7}$, $\Delta f_{10 \rightarrow 11}$, and $\Delta f_{14 \rightarrow 15}$. Black, red, magenta, and blue colors refer to multiple-replica simulations with $M = 1, 5, 10,$ and 15 , respectively. Dashed lines represent reference values calculated with MBAR method.³⁹

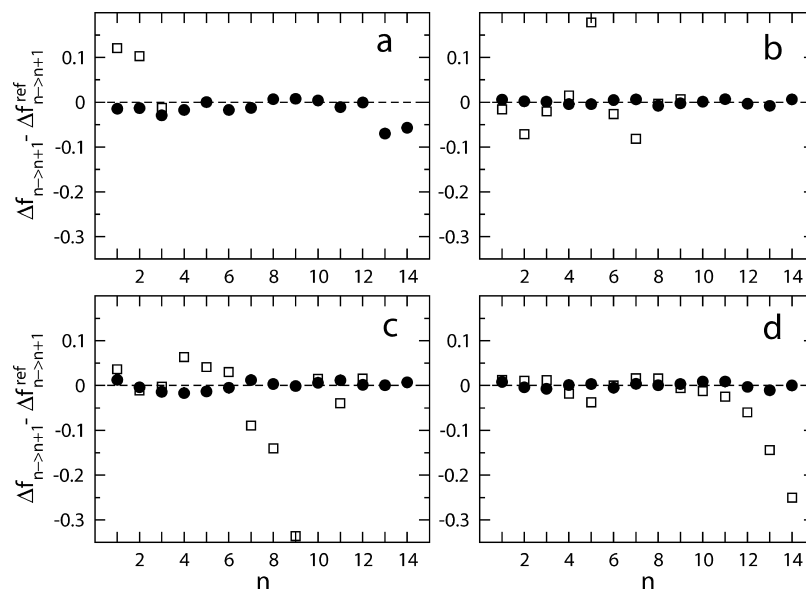


Figure 2. Differences $\Delta f_{n \rightarrow n+1} - \Delta f_{n \rightarrow n+1}^{\text{ref}}$ between BAR-SGE optimal weights, $\Delta f_{n \rightarrow n+1}$, and the reference ones, $\Delta f_{n \rightarrow n+1}^{\text{ref}}$ (from MBAR³⁹), as a function of n , computed from small-sample ST simulations. Panels a–d: $M = 1, 5, 10,$ and 15 . The values for two sampling times are reported (\square : 1.5×10^4 t -steps and \bullet : 1.5×10^6 t -steps). Dashed lines are drawn to highlight the zero.

potential energy employed in MBAR has been sampled with a frequency of 50 t -steps from 15 independent equilibrium simulations (one per ensemble/temperature) lasting 2.5×10^6 t -steps each (for a total of 7.5×10^5 configurations). The convergence of the MBAR optimal weights has been verified by calculations realized with an increasing number of analyzed configurations (the time-dependent MBAR optimal weights are available upon request). Hence, supported by the statistical sound, we may reasonably assume the MBAR weights as the “reference optimal weights”. Overall, it is encouraging that BAR-SGE weights converge to the reference ones already in the early stages of the simulations (note the scale on the ordinate axis in Figure 1), the number of replicas does not matter. In this respect, it is

important to consider that no initial guess for optimal weights is actually employed.

For a more global view of the data, in Figure 2 we report the difference $\Delta f_{n \rightarrow n+1} - \Delta f_{n \rightarrow n+1}^{\text{ref}}$ between BAR-SGE and MBAR optimal weights as a function of n . Specifically, we consider the differences obtained at the early stages and at the end of the simulations (up to 1.5×10^4 and 1.5×10^6 t -steps, respectively). For understanding the quantities into play, one should consider the large range of change of $\Delta f_{n \rightarrow n+1}^{\text{ref}}$, which goes from ~ 454 at $n = 14$ to ~ 503 at $n = 1$. For both times, $|\Delta f_{n \rightarrow n+1} - \Delta f_{n \rightarrow n+1}^{\text{ref}}|$ does not exceed 0.1% of $\Delta f_{n \rightarrow n+1}^{\text{ref}}$. In general, the performances of the algorithm increase with increasing the number of replicas, i.e., with improving the statistics, above all at short times. It is

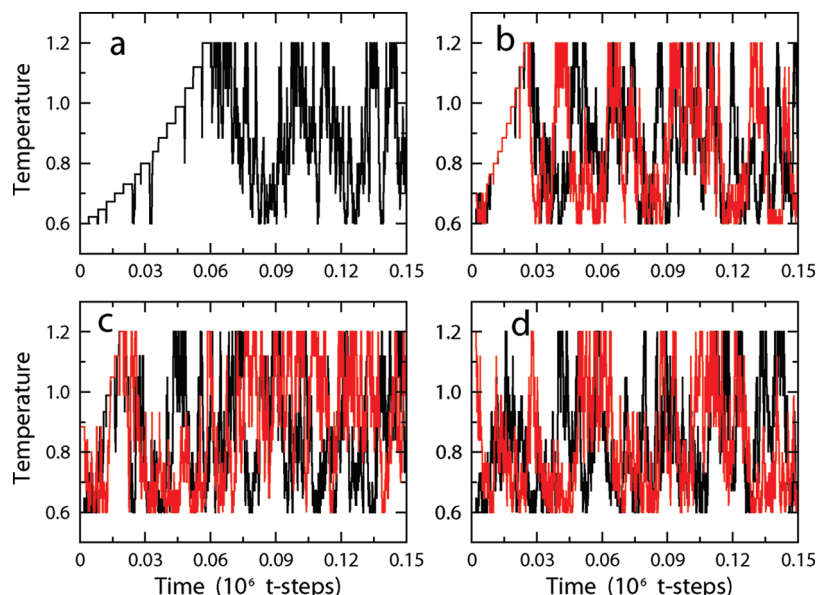


Figure 3. Temperature of selected replicas as a function of time per replica obtained from small-sample ST simulations. Panels a–d: $M = 1, 5$ (black and red replicas start from temperatures T_1 and T_5 , respectively), 10 (black and red replicas start from temperatures T_1 and T_{10} , respectively), and 15 (black and red replicas start from temperatures T_1 and T_{15} , respectively).

worthwhile observing the absence of several points in Figure 2 due to the fact that weight estimates are still not available. This occurs at the shortest time (1.5×10^4 t -steps), for large n and small M . Such a feature is explained considering that replicas are initially distributed in order of increasing temperature. This implies that the first available weight estimates are associated with transitions between ensembles at low temperature, corresponding to small n values. The remaining weights are obtained when ensembles at high temperature (large n values) start to be populated. In particular, for $M = 15$, optimal weights are available very soon because all ensembles are populated at the beginning of the simulation. This can be better appreciated in Figure 3, where we report the temperature of few replicas as a function of time per replica. In the single-replica simulation, a complete random walk in temperature is observable starting from about 6×10^4 t -steps. This time is reduced to 2×10^4 , 1.5×10^4 , and virtually, to 0 t -steps for $M = 5, 10$, and 15, respectively. An interesting feature observable in Figure 3 is the stair-like increase of the temperature in the initial part of the simulations. The step size is clearly correlated, but not necessarily equal, to the update frequency of optimal weights. After the highest temperature is reached, all replicas start to move through the ensembles with typical random walk. This can be observed for any M , though for large M , random walk may start well before the highest temperature ensemble is populated. This behavior highlights how the free energy perturbation approach (point 3b in Section 3.2) may enhance the exploration of ensembles in the early stages of the simulation.

It is also insightful to compare BAR-SGE method with other schemes, self-adaptive in principle, devised to update the optimal weights in SGE simulations. Recently, an interesting algorithm has been developed by Park, Ensign, and Pande³⁴ (ABWHAM) within the framework of Bayesian inference. ABWHAM is based on an updated scheme in which the information from previous data is stored in a prior

distribution, which is then updated to a posterior distribution according to the new data. The basic parameters of ABWHAM are the frequency of the histogram update (temperature histogram in ST and λ -histogram in a generic SGE simulation), the duration of the cycle of adaptation and sampling, the Ω factor which regulates the refresh of some variables of the method,³⁴ and most importantly, the initial guess for optimal weights. In our tests the temperature histogram is updated every 2 t -steps, while analysis is performed every 2000 t -steps. According to ref 34, we set $\Omega = 1$. No initial guess is actually used in ABWHAM, namely $f_n = 0$ for $n = 1, 2, \dots, 15$. Transitions between ensembles are attempted every 10 t -steps, while the simulation time is 5×10^6 t -steps per replica. We remark that our analysis is not aimed at establishing the superiority of one approach over the other (indeed, a systematic analysis on more complex systems would be needed) but rather to show how the choice of simulation parameters in the BAR-SGE method might not be as crucial for reaching convergence as it seems to be in the ABWHAM. The numerical comparison is shown in Figure 4. We observe that, while BAR-SGE algorithm gives accurate weights much before 5×10^5 t -steps (also see previous discussion), ABWHAM converges at very large times. In the latter method, we note a two-fold behavior. Noisy estimates are obtained up until a given threshold time, after which convergence is achieved in a very short period. This threshold time is variable and corresponds to the last refresh step.³⁴ The iterations before the last refresh step improve the initial guess and those after refine the posterior distribution. This feature was also observed in ST simulations of other simple models.³⁴ From Figure 4 we realize that statistical sampling is fundamental in reducing the threshold time. In fact, in the $M = 15$ simulation, it occurs at about 1.8×10^6 t -steps, while in the single-replica simulation, it is never reached during the whole simulation period. One could reduce the threshold time, and hence get a faster convergence, by increasing Ω .³⁴ A thorough analysis of this

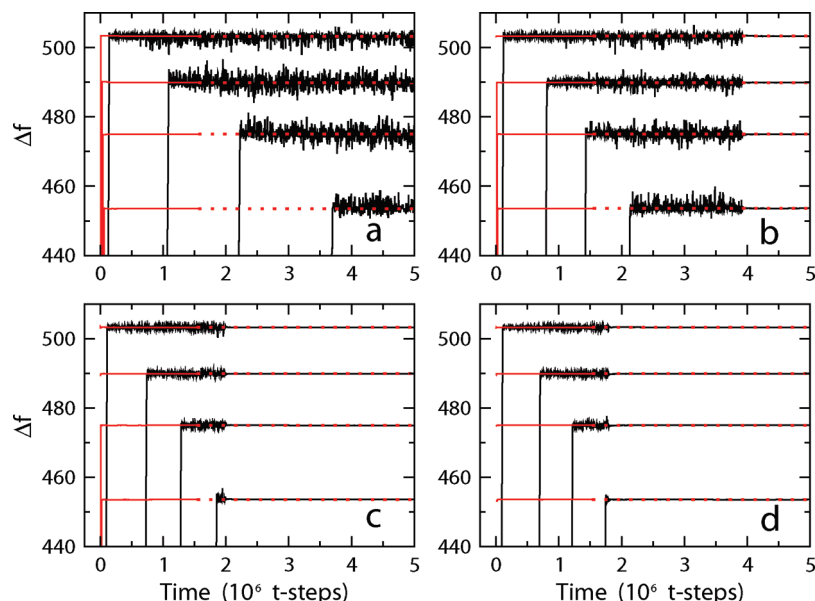


Figure 4. Comparison between BAR-SGE and ABWHAM optimal weights as a function of time per replica obtained from small-sample ST simulations (as in Figure 1). Red: BAR-SGE; black: ABWHAM. From top to bottom: Δf_{1-2} , Δf_{6-7} , Δf_{10-11} , and Δf_{14-15} . Panels a–d: $M = 1, 5, 10$, and 15 . Dotted lines represent extensions of the last-time weights calculated with BAR-SGE approach; they are drawn to make easier the comparison.

aspect would require a separate investigation and is far from the aim of the present work. Anyway, the most critical aspect of ABWHAM is the choice of the initial guess. If weights are comparable, then fast convergence can be achieved without initial guess.³⁴ However, when the optimal weights differ significantly, other methods, such as preliminary conventional simulations, are needed to obtain accurate initial guess and eventually to improve the convergence.²² De facto, this makes ABWHAM not fully self-consistent. On the other side, BAR-SGE algorithm allows to reach convergence without resorting to preliminary simulations. A good compromise between computational cost and convergence rate is roughly obtained when the number of replicas is comparable to the number of ensembles, a requirement that can be satisfied also with distributed computing clusters of modest size. Moreover, a positive fact is that the method is quite insensitive to the L_a and L_b parameters, provided N' is of the order of a few thousands. No significant differences are observed in convergence features by increasing N' (data not shown).

Concerning the computational cost of BAR-SGE, two important aspects must be remarked. First we note that no significant overhead is observed with respect to standard molecular dynamics simulations. The most time demanding task is the application of eq 22, which roughly takes a computer time comparable to that of a simulation step. From this point of view, ABWHAM is more efficient. However, since the update of the optimal weights is realized rarely, the overall elapsed times of BAR-SGE and ABWHAM simulations are comparable. Second, it is remarkable that, for a given simulation time per replica, the 5, 10, and 15 replica simulations are only 1.001, 1.002, and 1.003 slower than the single replica simulation.⁵⁵ These quite unexpected ratios come from two opposite effects. From one side, the use of many replicas/processors makes the simulation globally slower due to net communications between proces-

sors. From the other side, the simulation becomes faster because the sums of eqs 22–24 are distributed among the replicas/processors. Since the computational cost per replica is almost independent of the number of replicas/processors used in the simulation, we infer that the two competing effects are nearly balanced in our case. However, it is obvious that multiple replicas are preferable to single replica simulations if we want to enhance sampling for a given computer elapsed time.

5.1.2. Large-Sample Case. The biochemical systems typically investigated with molecular dynamics simulations are quite complex, not only because of the roughness of their free energy landscape but also due to the large number of degrees of freedom. Both aspects contribute to slow down the rate of convergence of any equilibrium sampling scheme, including generalized-ensemble methods. The complexity of the free energy landscape is intrinsically related to the kinetics of the sampling mechanisms, because strong structural rearrangements are often required. On the other side, the system size affects directly our capabilities of performing simulations long enough to produce adequate sampling. In ST simulations of large systems, an additional problem occurs. In order to get non-negligible acceptance ratios, a large number of ensembles/temperatures must be employed,⁵⁴ making the average transition rate between lowest and highest temperatures, and hence between free energy minima, slower. This is essentially due to the fact that the overlap of the potential energy distributions at two different temperatures decreases with increasing system size. Simulated solute tempering^{18,19} was just devised to reduce the number of atoms contributing to the potential energy distributions thus enhancing their overlap and eventually increasing the acceptance ratios. As a matter of fact, this could be a drawback when a SGE method, such as ABWHAM, is based on a thorough exploration of the temperature space. Moreover, it is unclear if ST simulations based on approximate estimates

Table 1. Reference Optimal Weights^a

n	$\Delta f_{n \rightarrow n+1}^{\text{ref}}$	n	$\Delta f_{n \rightarrow n+1}^{\text{ref}}$	n	$\Delta f_{n \rightarrow n+1}^{\text{ref}}$
1	3328.53	11	2536.52	21	1919.51
2	3240.23	12	2467.65	22	1865.84
3	3153.97	13	2400.49	23	1813.54
4	3069.99	14	2334.93	24	1762.54
5	2988.03	15	2270.98	25	1712.85
6	2907.98	16	2208.72	26	1664.35
7	2829.99	17	2147.93	27	1617.10
8	2753.77	18	2088.59	28	1571.02
9	2679.59	19	2030.86	29	1526.09
10	2607.14	20	1974.37		

^a Calculated from 30 independent equilibrium simulations using MBAR.³⁹ The temperatures are distributed following the rule $T_{n+1} = 1.02419 T_n$, where $T_1 = 0.6$.

of weight factors may yield effective sampling in the necessarily limited time of the simulation (think, e.g., to the replica exchange simulated tempering method³¹ or to the method based on potential energy averaging proposed in ref 22). In the present section, we address these issues by analyzing ST simulations of a medium–large sample (14 000 particles) realized with three sampling schemes, namely BAR-SGE, ABWHAM, and the standard method employing fixed weights obtained by averaging the potential energy from short preliminary simulations²² (from now on denoted with FW-SGE). In all cases, 30 ensembles/temperatures have been used ($N = 30$) with the temperature distribution rule reported in Section 4. To speed up the sampling, we have decided to use 30 replicas ($M = 30$), initially distributed over all ensembles (one replica per ensemble). The parameters for the BAR-SGE simulation are $L_a = 1$, $L_b = 1000$, $L_c = 10$, and $N' = 1000$. In the ABWHAM simulation, the temperature-histogram is updated every 1 t -step, while analysis is performed every 1000 t -steps. The other parameters of ABWHAM are those adopted in small-sample simulations. Reference optimal weights have also been calculated using MBAR.³⁹ Analogously to the small-sample case, in MBAR calculations the potential energy has been sampled with a frequency of 1 t -step from 30 independent equilibrium simulations (one per ensemble/temperature) lasting 5×10^5 t -steps each (five times longer than the ST simulations). The reference optimal weights, $\Delta f_{n \rightarrow n+1}^{\text{ref}}$, are reported in Table 1. The weight factors used in the FW-SGE simulations have been obtained following ref 22

$$g_{n+1} - g_n = \frac{1}{2}(\beta_{n+1} - \beta_n)(E_n + E_{n+1}) \quad (28)$$

for $n = 1, \dots, N - 1$. The quantities E_n and E_{n+1} are average potential energies estimated from standard simulations at the temperatures T_n and T_{n+1} . Here we report on the results of three FW-SGE simulations, indicated as FW-SGE-a, -b, and -c, whose weight factors are calculated by averaging the potential energy over 300, 1000, and 3000 t -steps, respectively. The deviations of the three sets of weight factors from the reference ones, $g_{n+1} - g_n - \Delta f_{n \rightarrow n+1}^{\text{ref}}$, are shown in Figure 5. We note that the absolute deviations are globally ordered as FW-SGE-a > -b > -c. This is simply due to the time interval considered for computing the average potential energies, which follows the reverse order. It is also important to note the almost systematic negative deviation of the

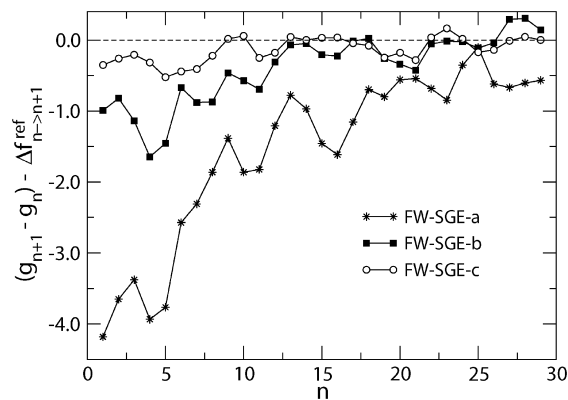


Figure 5. Deviations of the weight factors used in the fixed-weight ST simulations of the large sample from the reference ones calculated using MBAR³⁹ (the latter from Table 1). The weight factors have been calculated from 28, averaging the potential energy over 300 (FW-SGE-a: *), 1000 (FW-SGE-b: ■), and 3000 (FW-SGE-c: ○) t -steps in 30 standard simulations (one for each temperature). Dashed line represents the zero. Lines are drawn as a guide for eyes.

estimated weights from the reference ones, which is larger at lower temperature (small n values in Figure 5). This feature is clearly correlated with the fact that equilibrium is obtained in longer time at low temperatures. Lack of equilibrium is generally accompanied by an overestimate of the potential energy and, according to eq 28, by an underestimate of $g_{n+1} - g_n$. In spite of this, it is however worth noting that the weight factors of the FW-SGE-c simulation well approximate the reference ones, being the difference in most cases much lower than 0.5. Also the weight factors for the FW-SGE-b simulation approximate the reference weights quite satisfactorily, especially for temperatures higher than 0.744 (i.e., $n > 10$). Marked deviations from the ideal conditions are instead observed for the FW-SGE-a weights. In order to evaluate the efficiency of the average energy approach (summarized by eq 28) in producing random walks in temperature space, temperature histograms have been calculated from the FW-SGE-a, -b, and -c simulations. In particular, four histograms related to different time intervals are reported in Figure 6. The histograms obtained from a ST simulation performed with fixed optimal weights (those of Table 1) are also plotted for comparison (FW-SGE-ref in the figure). As expected, the FW-SGE-ref simulation yields almost flat histograms apart from the 0–25% time interval. Probably, in this case, the histogram keeps significant memory of the early stages of the simulation where equilibrium is still not attained. The features of the histograms computed from the FW-SGE-a, -b, and -c simulations are consistent with the estimated weight factors. The ensemble populations are inhomogeneous because weight factors deviate from the reference ones. Considering the (negative) deviations of $g_{n+1} - g_n$ (see Figure 5), we may also explain the large population of the low-temperature states. In fact, the apparent free energy difference between adjacent states, corresponding to $g_{n+1} - g_n$, is systematically smaller than the real (reference) value, $f_{n+1} - f_n$. As a consequence, the state with higher free energy, namely the $n + 1$ state, is sampled with a lower weight factor with respect to the ideal

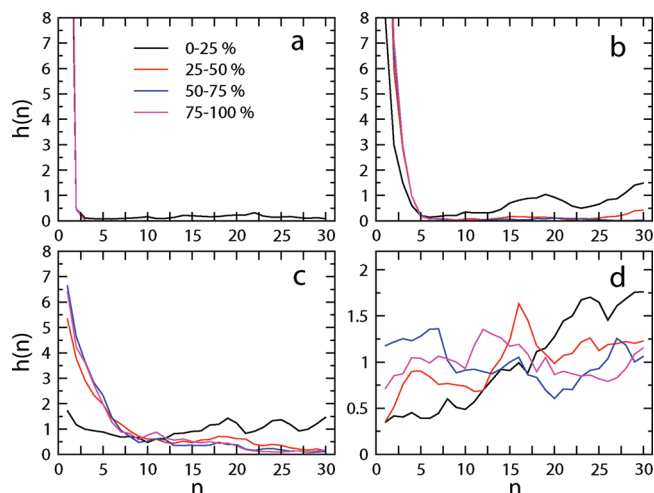


Figure 6. Ensemble/temperature populations as a function of the temperature label, n , computed from the fixed-weight ST simulations of the large sample. Panels a–d are FW-SGE-a, -b, -c, and -ref, respectively. The colors refer to the populations calculated in different time intervals (given as percentage of the total simulation time per replica).

case, ultimately leading to underpopulation of the state itself. A quite surprising aspect of the histograms of Figure 6 is instead the extent of inhomogeneity as compared to the observed deviations $g_{n+1} - g_n - \Delta f_{n \rightarrow n+1}^{\text{ref}}$. In the FW-SGE-a simulation, the population of states corresponding to $n > 2$ is practically 0. The flattening of the histograms slightly enhances passing to FW-SGE-b and then to FW-SGE-c simulations. However, also in the last case, although accurate weight factors are employed, the inhomogeneity remains significant. Note that the histograms observed in the 0–25% time interval keep strong memory of the initial homogeneous distribution of the replicas. The above observations suggest that, in order to get homogeneous sampling in ST simulations of large systems with fixed weight factors, temperature-dependent free energies (viz. weight factors) need to be estimated very accurately. Unluckily, adequate accuracy cannot be gained without efficient sampling. This vicious cycle supports the idea that only refinement protocols, such as BAR-SGE or ABWHAM, may ensure exhaustive sampling through the ensembles/temperatures. In the ABWHAM simulation reported here, the initial weight factors are those of the FW-SGE-b simulation, while no initial guess is employed for the BAR-SGE simulation. In Figure 7 we report the difference $\Delta f_{n \rightarrow n+1} - \Delta f_{n \rightarrow n+1}^{\text{ref}}$ between BAR-SGE/ABWHAM and MBAR optimal weights as a function of n (as resulting at the end of the simulations). The dispersion of $\Delta f_{n \rightarrow n+1} - \Delta f_{n \rightarrow n+1}^{\text{ref}}$ about the zero obtained from ABWHAM is due to the occurrence of refresh steps (see also discussion in Section 5.1.1). However, although full convergence is not reached with ABWHAM, the weights calculated by averaging the estimates over the whole simulation run provide much better agreement with the reference (see asterisks in Figure 7). The optimal weights estimated from BAR-SGE are instead very accurate. These convergence features are pretty mirrored by the temperature histograms obtained from the two methods (see Figure 8). The flattening of the histogram during the progress of the

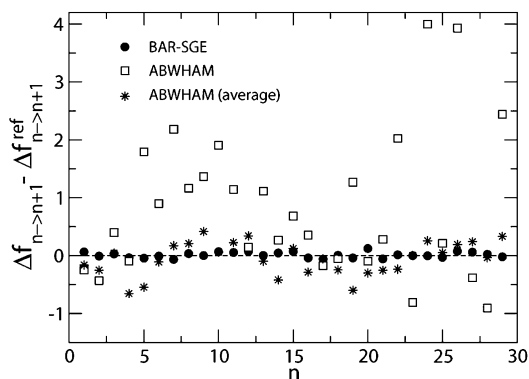


Figure 7. Differences $\Delta f_{n \rightarrow n+1} - \Delta f_{n \rightarrow n+1}^{\text{ref}}$ between BAR-SGE/ABWHAM optimal weights, $\Delta f_{n \rightarrow n+1}$, and the reference ones, $\Delta f_{n \rightarrow n+1}^{\text{ref}}$ (from MBAR³⁹) as a function of n , computed from large-sample ST simulations. The full circles (●) indicate the differences of the BAR-SGE estimates. The open squares (□) indicate the differences of the ABWHAM estimates performed at the last simulation step. The asterisks (*) indicate the differences calculated by averaging the ABWHAM estimates done at each analysis (see text for details). Dashed line is drawn to highlight the zero.

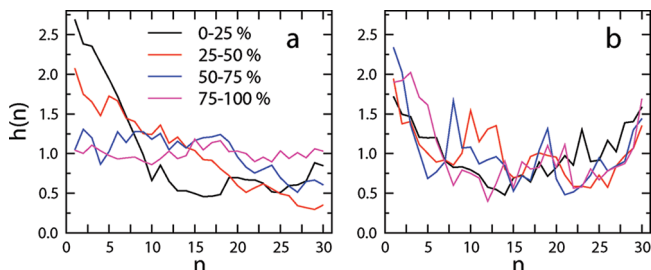


Figure 8. Ensemble/temperature populations as a function of the temperature label, n , computed from the BAR-SGE/ABWHAM ST simulations of the large sample. Panels a and b are BAR-SGE and ABWHAM, respectively. The colors refer to the populations calculated in different time intervals (given as percentage of the total simulation time per replica).

simulation is more evident for BAR-SGE than for ABWHAM, consistently with the noisy trend of the ABWHAM weights. Finally, it is remarkable that in the last time interval (75–100%), BAR-SGE and FW-SGE-ref give comparable results.

In BAR-SGE, the refinement of the optimal weights, $\Delta f_{n \rightarrow n+1}$ (for $n = 1, \dots, N - 1$), is based on the periodic estimate of free energy uncertainties (eq 23), employed in the weighted average of eq 25 (see Section 3.3). For each $\Delta f_{n \rightarrow n+1}$, the set of uncertainties calculated during the simulation provides also the global error, $\delta(\Delta f_{n \rightarrow n+1})$, via eq 26. In the present case, all $\delta(\Delta f_{n \rightarrow n+1})$ fall in the range 0.0077–0.0105, the average value being 0.0092. The errors on the optimal weights can give information about the probabilities of visiting the various ensembles/temperatures. We know that, if $\Delta f_{n \rightarrow n+1}$ were not affected by error, then all ensembles/temperatures would be populated with the same probability. In such a situation, the ratio between the probabilities of two ensembles, say n and m , can be written as $P_n/P_m = Z_n/Z_m \exp(\Delta f_{m \rightarrow n}) = 1$ (see eq 9). If $\Delta f_{m \rightarrow n}$ is

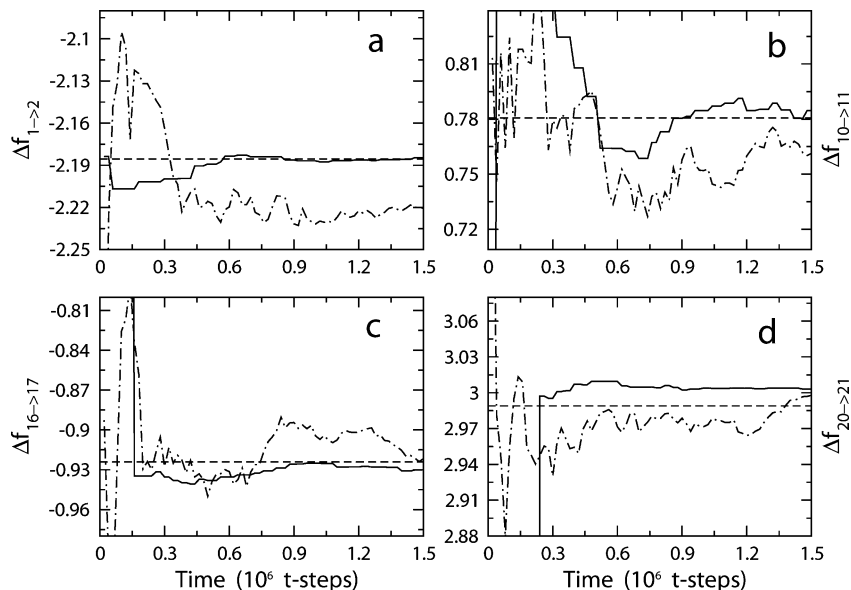


Figure 9. Representative optimal weights as a function of time per replica obtained in 10-replica SGE simulations in λ -space. Panels a–d: $\Delta f_{1\rightarrow 2}$, $\Delta f_{10\rightarrow 11}$, $\Delta f_{16\rightarrow 17}$, and $\Delta f_{20\rightarrow 21}$. Solid and dot-dashed lines are obtained from simulations using BAR-SGE scheme and ABWHAM, respectively. Dashed lines represent reference values calculated by thermodynamic integration.

affected by the error $\delta(\Delta f_{m\rightarrow n})$, then the ratio P_n/P_m will vary the most in the range:

$$\frac{P_n}{P_m} = \frac{Z_n}{Z_m} e^{\Delta f_{m\rightarrow n} \pm \delta(\Delta f_{m\rightarrow n})} = e^{\pm \delta(\Delta f_{m\rightarrow n})} \quad (29)$$

In the previous equation the error coming from histogram sampling has been assumed negligible. Therefore, it represents the error associated with inaccurate determination of the optimal weights rather than with inaccurate sampling of the temperature space. From eq 29 we infer that errors in determining optimal weights do affect the ratio in asymmetric way. Symmetry is obtained in the limit of small $\delta(\Delta f_{m\rightarrow n})$ (expand the exponential of eq 29 in Taylor's series about the zero). Considering the maximum error on $\Delta f_{n\rightarrow n+1}$ in our simulation, i.e. 0.0105, the previous equation establishes that the ratio P_n/P_{n+1} ranges in the interval 0.99–1.01 (difference of $\sim 1\%$ with respect to the theoretical value of 1). An overestimate of the maximum change in the ratio P_N/P_1 involving the end states can also be gained from eq 29 assuming that

$$\delta(\Delta f_{1\rightarrow N}) = \sum_{n=1}^{N-1} \delta(\Delta f_{n\rightarrow n+1}) \quad (30)$$

We have found $P_N/P_1 = 0.77\text{--}1.31$, which corresponds to a maximum deviation from 1 by 31%.

5.2. SGE Simulations in λ -Space. As previously stated, we also report on the results of a SGE simulation performed in ensembles associated with a parameter, λ , bound to the distance between two particles (λ -ensembles). Although various SGE simulations have been carried out ($M = 1, 5$, and 10), we decided to report only the outcomes of the 10-replica simulation, because the features dependent on M are similar to those discussed for ST simulations. The relevant parameters in t -step units are $L_a = 10$, $L_b = 2 \times 10^4$, $L_c = 100$, and $N' = 2000$. The convergence features of the method

are shown in Figure 9, where we report four representative optimal weights corresponding to the ensemble transitions $\lambda_1 = 0.5 \rightleftharpoons \lambda_2 = 0.65$, $\lambda_{10} = 1.85 \rightleftharpoons \lambda_{11} = 2.0$, $\lambda_{16} = 2.75 \rightleftharpoons \lambda_{17} = 2.9$, and $\lambda_{20} = 3.35 \rightleftharpoons \lambda_{21} = 3.5$. Results from a 10-replica simulation using ABWHAM are also reported in the figure for comparison. In this last simulation, the λ -histogram is updated every 10 t -steps, while weight analysis is performed every 2×10^4 t -steps. As usual, $\Omega = 1$. Transitions between ensembles are attempted every 100 t -steps, while the simulation time is 1.5×10^6 t -steps per replica. At variance with the ST case, in this ABWHAM simulation we have used an initial guess for optimal weights, drawn from a prior ABWHAM-based simulation of 1.5×10^6 t -steps per replica, during which refresh was active. Note that, in the present simulation, no refresh steps were necessary. Reference optimal weights from thermodynamic integration²³ are also plotted in Figure 9. Thermodynamic integration data are recovered from canonical simulations of 5×10^6 t -steps (density = 0.85 and temperature = 0.6). The dimensionless Hamiltonian associated with the various ensembles is reported in eq 15, with a force constant k of 25. The λ -step size for numerical integration is 0.05. From Figure 9 we note that the two update methods give comparable convergence. We must, however, remember that ABWHAM weights come from a longer simulation history targeted to the initial guess. It is remarkable that, in the BAR-SGE method, even the early estimates well agree with the values obtained from thermodynamic integration and from ABWHAM. Comparable to ST simulations, λ -ensembles are populated very quickly. This is clearly shown in Figure 10, where we report λ as a function of time per replica. The features of Figure 10 strongly resemble those of Figure 3, whether in the random walk through the various ensembles or in the stair-like trend characterizing the λ evolution at early times.

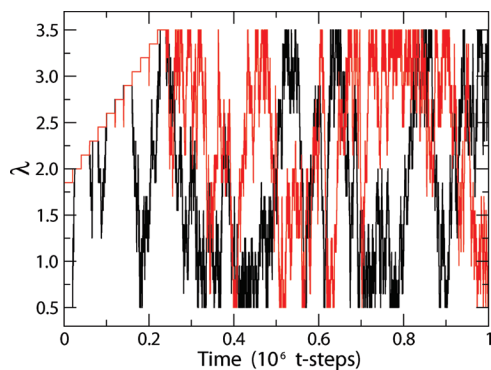


Figure 10. Value of λ as a function of time per replica for two replicas taken from the BAR-SGE-based 10-replica simulation. Black and red lines are related to replicas starting from λ_1 and λ_{10} , respectively.

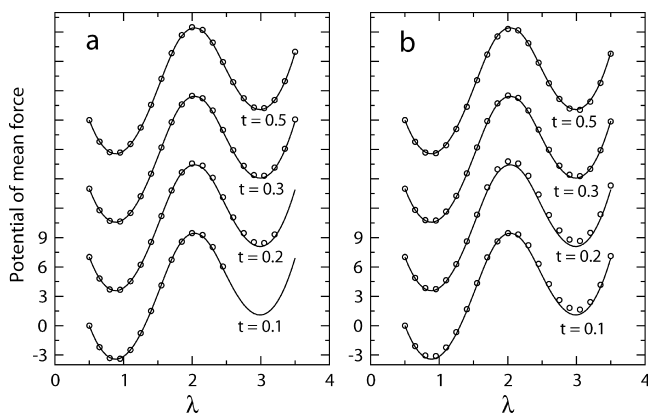


Figure 11. Potential of mean force (adimensional units) as a function of λ calculated from 10-replica SGE simulations at various times (in 10^6 units). Open circles (\circ): data from SGE simulations, and solid lines: data from thermodynamic integration. Panels a and b: simulations adopting BAR-SGE scheme and ABWHAM, respectively. For the sake of clarity PMF profiles are shifted.

Furthermore, it is instructive to analyze how the PMF along the λ coordinate is built up during the sampling. The PMF is recovered from the optimal weights as

$$f(\lambda_n) = \sum_{i=1}^{n-1} \Delta f_{i \rightarrow i+1} \quad (31)$$

In Figure 11, we plot the PMF calculated at various times with BAR-SGE and ABWHAM approaches and compare such profiles to the reference one. The most evident feature is that BAR-SGE method, at variance with ABWHAM, needs a certain time to complete PMF construction. This time may depend on the system type and, in general, can be reduced by increasing the number of walking replicas (see discussion in Section 5.1). On the other side, the PMF curve at early times (see $t = 0.1$ curve in Figure 11a), although incomplete, is very accurate and would not seem to require further refinement. However, for better evaluating the relative (though not optimized) performances of the BAR-SGE scheme and the ABWHAM, we must remember that in the latter case a preliminary simulation has been carried out to recover an initial guess. We finally note that the errors on

the free energy differences between adjacent states calculated by eq 26 fall well below 0.01. The maximum error on the free energy difference between the end states calculated from eq 30 is 0.07.

Unbiased PMF profiles along the collective coordinate associated with λ (the interparticle distance in our case) can also be calculated in posterior analysis (data not shown) using multiple-histogram reweighting techniques^{25,26} or other recent approaches developed in the framework of nonequilibrium statistical mechanics.^{46,47}

5.3. How Eq 24 Does Affect the Acceptance Ratio in SGE Simulations. The effect of using eq 24 in SGE simulations is that of enhancing the acceptance ratio for those transitions that promote a replica toward ensembles that have not been visited. Suppose, for instance, to set up a M -replica ST simulation with N ensembles (with $N > M$) by associating replica 1 to the ensemble with temperature T_1 , replica 2 to the ensemble with temperature T_2 , and so on, until the ensemble with temperature T_M . As usual, we assume that the temperatures are in order of increasing index and that transitions occur only between neighboring temperatures. On the basis of the BAR-SGE scheme, the transition $T_M \rightarrow T_{M+1}$ can be attempted only using an estimate of $\Delta f_{M \rightarrow M+1}$ from eq 24. In fact, works $W[M+1 \rightarrow M]$, needed to employ eq 22, are not available because the ensemble $M+1$ has never been visited. A similar situation would occur if the replicas were distributed with reverse order. Therefore, the free energy estimates provided by eq 24 are important in the early stages of the simulation because they affect directly the diffusion of replicas through the ensembles.

As an example, we calculate the distribution function of the acceptance ratio for the transition $T_{13} \rightarrow T_{14}$ in our model system. To this aim, we consider all $W[13 \rightarrow 14]$ work values recorded during the 15-replica ST simulation. For our purpose, since we are interested only in a set of work values, M does not matter. Then we have partitioned the set of works in several independent subsets, each made of D elements (here $D = 100, 300,$ and 1000). For each subset we have calculated $\Delta f_{13 \rightarrow 14}$ according to eq 24, thus obtaining a collection of reliable optimal weights. These weights have then been employed to compute the average acceptance ratio from the whole original set of works. In such a way it is possible to construct distribution functions of average acceptance ratios. The distribution functions recovered using $D = 100, 300,$ and 1000 are plotted in Figure 12. They are very broad, but the relevant fact is the shift toward higher values of the acceptance ratio with decreasing D , namely the number of work samples used for calculating $\Delta f_{13 \rightarrow 14}$. The average acceptance ratio is 0.31, 0.27, and 0.24 for $D = 100, 300,$ and 1000 , respectively. These differences arise from the fact that $\Delta f_{13 \rightarrow 14}$ is as much overestimated as D is smaller, in agreement with previous observations on the convergence properties of work exponential averages.⁵² When D increases, $\Delta f_{13 \rightarrow 14}$ approaches the exact value as well as the resulting acceptance ratio. This conclusion is also supported from the average acceptance ratio obtained using the reference optimal weight (from MBAR). Its value, 0.22, is $\sim 9\%$ smaller than that obtained from 1000 samples. This difference, though not negligible, reveals that already 1000

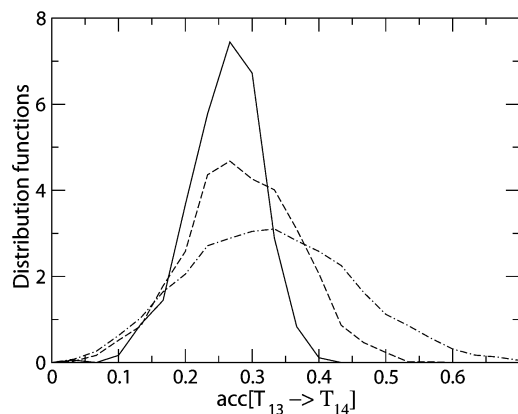


Figure 12. Normalized distribution functions of the average acceptance ratio for the transition $T_{13} \rightarrow T_{14}$. Solid, dashed, and dot-dashed lines are related to the distributions calculated by using $D = 1000, 300,$ and $100,$ respectively.

samples are sufficient to get good free energy estimates from eq 24. In fact, the reference value of $\Delta f_{13 \rightarrow 14}$, 459.79, is only 0.12 smaller than the average value calculated using $D = 1000$.

6. Concluding Remarks

In serial generalized-ensemble simulations, such as simulated tempering, weight factors must be determined somehow to allow a random walk in the space of the chosen collective coordinate (the temperature in simulated tempering). In this respect, adaptive methods, such as BAR-serial generalized-ensemble (BAR-SGE) and Bayesian weighted histogram analysis method (ABWHAM), may provide effective routes to the fast determination of weight factors without resorting to preliminary simulations. This is indeed an advantageous feature of BAR-SGE and ABWHAM because, as we have shown in the present work (Section 5.1.2), initial estimates of weight factors from preliminary simulations must be very accurate to ensure an almost random walk of the replicas through the ensemble space. Even a small underestimate of the weight factors, which typically occurs as equilibrium is still not achieved, may lead to significant inhomogeneous sampling. In this respect, the BAR-SGE method offers interesting perspectives in enhancing the convergence of optimal weights with minimal introduction of tunable parameters. The truly relevant parameter entering into play is the update frequency of weights, which must ensure the storage of a sufficient number of work samples (see eq 20) needed to get accurate free energy estimates (see eq 22). The minimum value of the number of samples ranges from one thousand to a few thousand. It is also important to remark that in a suitable adaptive method, each update should in principle account for the uncertainty associated with the individual estimates. BAR-SGE scheme includes such a feature by a variance-weighted sum of the individual estimates (Section 3.3). In SGE simulations realized in the space of a collective coordinate of the system, the possibility of calculating the uncertainties of the free energy differences between neighboring ensembles provides a way of estimating the error in the potential of mean force. Furthermore, since the update of a single weight involves data from only two

neighboring ensembles, the computational cost of BAR-SGE is much smaller than that of multiple-histogram reweighting. In the case of our BAR-SGE simulations, using 5, 10, and 15 replicas leads to an increase of the elapsed time per replica by only 1.001, 1.002, and 1.003, respectively, with respect to the single-replica simulation. This put forward the BAR-SGE algorithm as a suitable methodology for large computing distributed environments.

Acknowledgment. The author is grateful to Gianfranco Lauria (LENS, University of Florence, Italy) for technical support in computer facilities at LENS and to Simone Marsili (Department of Chemistry, University of Florence, Italy) for providing a program for MBAR calculations. This work was supported by European Union Contract RII3-CT-2003-506350.

References

- (1) Okamoto, Y. *J. Mol. Graphics Modell.* **2004**, *22*, 425.
- (2) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249.
- (3) Berg, B. A. *Int. J. Mod. Phys. C* **1992**, *3*, 1083.
- (4) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451.
- (5) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776.
- (6) Rauscher, S.; Neale, C.; Pomes, R. *J. Chem. Theory Comput.* **2009**, *5*, 2640.
- (7) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140.
- (8) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
- (9) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604.
- (10) Tesi, M. C.; van Rensburg, E. J. J.; Orlandini, E.; Whittington, S. G. *J. Stat. Phys.* **1996**, *82*, 155.
- (11) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96.
- (12) Mitsutake, A.; Okamoto, Y. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2009**, *79*, 047701.
- (13) Lee, A. J.; Rick, S. W. *J. Chem. Phys.* **2009**, *131*, 174113.
- (14) Ballard, A. J.; Jarzynski, C. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12224.
- (15) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *J. Chem. Phys.* **2003**, *118*, 6664.
- (16) Mitsutake, A.; Okamoto, Y. *J. Chem. Phys.* **2004**, *121*, 2491.
- (17) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703.
- (18) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13749.
- (19) Denschlag, R.; Lingenheil, M.; Tavan, P.; Mathias, G. *J. Chem. Theory Comput.* **2009**, *5*, 2847.
- (20) Escobedo, F. A.; Martinez-Veracoechea, F. J. *J. Chem. Phys.* **2008**, *129*, 154107.
- (21) Trebst, S.; Huse, D. A.; Troyer, M. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2004**, *70*, 046701.
- (22) Park, S.; Pande, V. S. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2007**, *76*, 016703.
- (23) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.

- (24) McQuarrie, D. A. *Statistical Mechanics*; HarperCollins Publishers: New York, 1976.
- (25) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.
- (26) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195.
- (27) Park, S. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2008**, *77*, 016709.
- (28) Zhang, C.; Ma, J. *J. Chem. Phys.* **2008**, *129*, 134112.
- (29) Hansmann, U. H. E.; Okamoto, Y. *J. Comput. Chem.* **1997**, *18*, 920.
- (30) Irbäck, A.; Potthast, F. *J. Chem. Phys.* **1995**, *103*, 10298.
- (31) Mitsutake, A.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *332*, 131.
- (32) Huang, X.; Bowman, G. R.; Pande, V. S. *J. Chem. Phys.* **2008**, *128*, 205106.
- (33) Zhang, C.; Ma, J. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2007**, *76*, 036708.
- (34) Park, S.; Ensign, D. L.; Pande, V. S. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2006**, *74*, 066703.
- (35) Chelli, R.; Marsili, S.; Barducci, A.; Procacci, P. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2007**, *75*, 050101.
- (36) Chelli, R. *J. Chem. Phys.* **2009**, *130*, 054102.
- (37) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245.
- (38) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.
- (39) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
- (40) In Monte Carlo generalized-ensemble simulations, momenta are dropped out.
- (41) Mitsutake, A.; Okamoto, Y. *J. Chem. Phys.* **2009**, *130*, 214105.
- (42) Here, we assume implicitly that the indexes n and m belong to an ordered list such that $T_1 < T_2 < \dots < T_N$ or $\lambda_1 < \lambda_2 < \dots < \lambda_N$.
- (43) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695.
- (44) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1986**, *34*, 2499.
- (45) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635.
- (46) Minh, D. D. L.; Adib, A. B. *Phys. Rev. Lett.* **2008**, *100*, 180602.
- (47) Nicolini, P.; Procacci, P.; Chelli, R. *J. Phys. Chem. B* **2010**, in press.
- (48) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. *Phys. Rev. Lett.* **2003**, *91*, 140601.
- (49) Williams, S. R.; Searles, D. J.; Evans, D. J. *Phys. Rev. Lett.* **2008**, *100*, 250601.
- (50) Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177.
- (51) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690.
- (52) Gore, J.; Ritort, F.; Bustamante, C. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12564.
- (53) Cowan, G. *Statistical Data Analysis*; Oxford University Press: Oxford, U.K., 1998.
- (54) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058.
- (55) The calculations were performed on a distributed computing cluster made of personal computers communicating each other with maximum net speed of 2 Gb s⁻¹.

CT100105Z

JCTC

Journal of Chemical Theory and Computation

Assessment of DFT and DFT-D for Potential Energy Surfaces of Rare Gas Trimers—Implementation and Analysis of Functionals and Extrapolation Procedures

Roberto Peverati, Marina Macrina, and Kim K. Baldridge*

University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received February 2, 2010

Abstract: Given the recent developments in methodology associated with the accurate computation of molecular systems with weak interactions, it is of particular interest to revisit systems that are notoriously challenging for determining reliable potential energy surface (PES) descriptions. Additionally, challenges associated with carrying out complete basis set extrapolation procedures and treatment of basis set superposition error (BSSE) are of importance in these descriptions. In this work, investigation into the ability to accurately predict the potential energy surfaces of the main Rg₃ molecules (Rg = He, Ne, Ar) is made across a range of wave function types and large basis sets, including the use of several established extrapolation procedures and counterpoise corrections. Wave function types span most classes of density functional types, including the newest DFT-D schemes, and are benchmarked against high accuracy CCSD(T)/CBS methodology. Study of such systems is valuable, as they serve as simple models for many complex properties, most importantly *n*-body weak interaction energies.

Introduction

Rare gas compounds are, in many cases, simple models for the study of complex properties. In particular, weak interactions of van der Waals bound dimers, *n*-body interaction energies in trimers, tetramers, etc., and complex aggregations of large clusters, are of significant importance. Few-body, rare gas compounds are heavily used for parametrizations of semiempirical potentials, for example, in empirical force fields or *ab initio* molecular dynamics methods. In these cases, highly accurate *ab initio* potential energy surfaces (PES) are extremely important for such parametrizations. Also significant are three-body atomic systems, which present intriguing properties, such as “Efimov physics”,^{1–4} and “Borromean states”,^{5,6} with the presence of bound states even when the analogous two-body systems are unbound.

Homodimers present a well-known, simple, one-dimensional PES,⁷ commonly used in parametrizations. Equilateral homotrimers present an analogous one-dimensional PES but with a much larger (Borromean) bound state. Because of a less profound knowledge of the nature of the bonding, their

use in parametrizations is restricted mainly to three-body correction components.^{8,9} For similar reasons, while accurate results of rare gas dimers are used as test systems for theoretical models, homotrimers are very rarely used in this sense.

The study of rare gas trimers is nevertheless a very well established first step in the investigation of the stability of large clusters, and both experimental and theoretical studies have contributed data in this direction.^{10–20} Several experimental studies have been carried out on rare gas triatomic systems,²¹ indicating less sensitivity in measurements than the weaker binding dimers. There is very little in the literature detailing accurate calculations of the potential energy surfaces of the rare gas trimers, however. Additionally, none of the more recent dispersion-enabled DFT functionals have been tested on these systems, despite their relevance for understanding intermolecular interactions and the implications on the overall computational cost savings compared to traditional benchmark level methods. Rare gas trimers represent ideal candidates for parametrization and validation of new theoretical models, in the same way that dimers have been used up to now.

* Corresponding author tel.: +41 44 635 4201; fax: +41 44 635 6888; e-mail: kimb@oci.uzh.ch.

Table 1. Strategies for Density Functionals beyond GGA

density functional type	main feature	examples	reference
meta-GGA	depend on the Kohn–Sham kinetic energy density	MOX family τ -HCTH family BMK	Zhao et al. ^{22–25} Handy and Boese ²⁶ Boese/Martin ²⁷
range-separated hybrid (RSH)	Coulomb operator is separated into long-range and short-range terms, the extent of which determines the exact variant of the functional	LC-BLYP, CAM-B3LYP HSE ω B97	Savin et al. ^{28,29} Scuseria and Heyd ^{30,31} Chai and Head-Gordon ³²
empirical/semiempirical	vdW dispersion interactions described empirically with a damped interatomic R^{-6} potential	B97-D	Grimme ³³
double-hybrid	includes terms derived from correlated wave function methods (e.g., MP2 theory)	B2-PLYP, mPW2-PLYP MC3BB, MC3MPW B2K-PLYP, mPW2K-PLYP	Grimme and Schwabe ^{34,35} Zhao et al. ³⁶ Martin et al. ³⁷
Andersson–Langreth–Lundqvist functional	long-range exchange correction scheme together with the Andersson–Langreth–Lundqvist vdW functional	vdW-DF	Langreth et al. ³⁸

There are now a growing number of theoretical models that are appropriate for treating the structure and properties of weakly bound clusters, enabling greater understanding of the importance and representation of short-, intermediate-, and long-range interaction, particularly since the introduction of entire new generations of approximations for the exchange-correlation potential (e.g., see Table 1 and references therein). The local density approximation (LDA) and its analogue, local spin density approximation (LSDA), as the first approximations to the exchange-correlation potential ν_{xc} used by Kohn and co-workers with high success, were particularly applicable in solid-state physics. Initial strategies to improve LDA/LSDA enhanced the exchange-correlation functional terms that depend on the gradient of the density, leading to generalized gradient approximation (GGA) functionals. Despite their great success, GGA approximations fail in the description of properties that depend mainly on the correlation of electrons, such as is the case for rare gas trimer complexes. To overcome the main limitations of GGA functionals, various strategies have been developed (Table 1), which offer a high degree of reliability in these cases.

In this work, we have evaluated the latest dispersion sensitive density functional theory (DFT) based methodologies for their ability to accurately represent the potential energy surfaces for a series of rare gas trimer systems. After establishing MP2, CCSD, and CCSD(T) benchmark potentials, a general assessment of DFT trimer potential energy surfaces is made. Two technical issues concerning grid size and BSSE are addressed, before concluding remarks.

Theoretical Methodology and Approach

All calculations reported here used a locally modified version of the GAMESS electronic structure program, running on our group cluster hardware. Associated visualization and analysis was carried out using MacMolPlt³⁹ and Qutemol.⁴⁰ In the present work, we apply our recently implemented semiempirically corrected density functionals,⁴¹ in addition to key functionals implemented by several other GAMESS DFT contributors. Moreover, to carry out a full analysis on the series of rare gas trimers across various classes of density

functional types, we have implemented a large variety of additional functionals of different class types. In the process, we have facilitated testing of parameters, future implementations of new functionals, updates of existing functionals, and, from the earlier work, the ability to include the semiempirical dispersion correction in various functionals. Initially, two new routines for the calculation of the B97 family of functionals^{42–45} were implemented. The B97-D functional is a special reparameterization of the original Becke 1997 functional,⁴² produced by Grimme³³ with the purpose of avoiding the effect of double-counting in the vdW region. The power expansion series coefficients of the original functional description were optimized by Grimme to restrict the density functional representation of the shorter electron correlation ranges, while the medium- to long-range representation is handled by the semiempirical correction term.

Following the formulation of the original B97 functional,⁴² we separate exchange and correlation contributions. For the correlation, a FORTRAN routine was generated via a modified version of the *dfauto* program of Knowles et al.⁴⁶ The expansion in the gradient up to five terms was used, and the numerical coefficients were passed to the routine as parameters. This correlation routine enables calculation of correlation energy for all implemented functionals. The exchange component of the B97 functional has been implemented as a separate routine, composed of three distinct blocks, (1) the LSDA component, implemented using the previous LSDA routine of GAMESS that includes the range-separation of the Coulomb operator (for the ω B97 family of functionals), (2) the GGA component, implemented using the modified *dfauto* program, as done for the correlation component, and (3) the τ -dependency for the τ -HCTH family of functionals. All three components are accumulated together appropriately, and the global functional derivatives are calculated using a simple chain-rule formula. The new routines enable the implementation of a large number of different reparameterizations of the same basic functional form, and because of their modular nature, new sets of parameters can be easily added in the future, either for the purpose of refinement of the existing formulations based on

new data or for the formulation of new functionals. In addition, our recent implementation of semiempirical dispersion correction capabilities⁴¹ can also be readily accessed for the functionals, enabling, for example, the B97-D and the ω B97X-D dispersion corrected forms.

Several functionals are hybrid functionals, with the percentage of HF exchange appropriately added via an array value option. The range-separated HF and DFT exchange of the RSH functionals are calculated with the long-range correction scheme of GAMESS, but using a slightly modified routine to allow the multiplication with the GGA correction. The LR-HF integrals are calculated directly using Savin et al.'s operator⁴⁷ in the two-electron integrals module of GAMESS. According to the formula of Head-Gordon and Chai,³² the ω B97X functional is calculated as

$$E_{XC}^{LC-GGA} = E_X^{SR-GGA} + E_C^{GGA} + E_X^{LR-HF} + c_x E_X^{SR-HF} \quad (1)$$

The scaled SR-HF exchange for the ω B97X functional is obtained in our implementation indirectly by using the factorization of the total HF exchange (calculated without Savin et al.'s modified Coulomb operators) as

$$c_x E_X^{SR-HF} = c_x (E_X^{HF} - E_X^{LR-HF}) \quad (2)$$

The ω B97X functional is then implemented in a slightly unconventional way, as

$$E_{XC}^{LC-GGA} = E_X^{SR-GGA} + E_C^{GGA} + (1 - c_x) E_X^{LR-HF} + c_x E_X^{HF} \quad (3)$$

This particular reformulation of the functional definition of eq 1 provides additional insight, since it shows more clearly the similarities between the range separated hybrids and the simple meta-GGA functional form of τ -HCTH, as well as other derived functionals.

$$E_{XC}^{meta-GGA} = E_X^{GGA} + E_C^{GGA} + E_X^\tau + c_x E_X^{HF} \quad (4)$$

In eq 3, in fact, the τ -dependent term of eq 4 is substituted by the long-range Hartree–Fock term (nonlocal by definition), and the GGA exchange is limited to short-range in eq 3. The other terms, the GGA correlation and the scaled HF exchange, are exactly the same for both functional forms. Written in this form, it is also clear that ω B97 is closely related to nonhybrid functionals ($c_x = 0$), while ω B97X is related to hybrid functionals ($c_x \neq 0$). A summary of all B97-related functionals available in the new release of GAMESS through the described new routines can be found in the Supporting Information.

Evaluated functionals for the present work include **B97**,⁴² **B97-1**,⁴³ **B97-2**,⁴⁵ **B97-3**,⁴⁴ **B97-D**,³³ **B98**,⁴⁸ **HCTH/93**,⁴³ **HCTH/120** and **HCTH/147**,⁴⁹ **HCTH/407**,⁵⁰ τ -**HCTH**,²⁶ τ -**HCTHhyb**,²⁶ **BMK**,²⁷ ω **B97** and ω **B97X**,³² ω **B97X-D**,³² **BLYP**,^{51–53} **B3LYP**,^{22,54,55} **B2-PLYP**,³⁴ **CAM-B3LYP**,²⁹ **VS98**,⁵⁶ **PKZB**,⁵⁷ **TPSS**,^{58,59} **TPSSH**,^{60,61} **TPSSM**,⁶² **M05**,²² **M05-2X**,⁶³ **M06** and **M06-2X**,²⁴ **M06-L**,⁶⁴ **M06-HF**,⁶⁵ and **M08-HX** and **M08-SO**.²⁵ The functionals in bold are our most recently implemented functionals to the GAMESS suite. All computations have been carried out using the (96, 1202) Lebedev grid⁶⁶ (called the ‘army’ grid in GAMESS), and

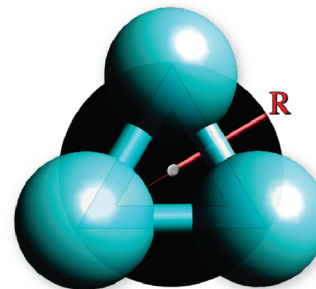


Figure 1. Radial coordinate of the Rg_3 D_{3h} trimer molecule.

an additional set using the (400, 770) Lebedev grid. A general grid convergence investigation was carried out for the meta-GGA functionals using several other grid specifications as detailed in the text. In accord with our previous study on the performance of the B97-D functional,⁴¹ the scaling factor for the semiempirical dispersion was taken as $s_6 = 1.00$ for all the double- ζ basis sets considered and $s_6 = 1.25$ for all the triple- ζ basis sets considered. Basis set superposition error (BSSE) is corrected with the counterpoise (CP) method.⁶⁷ A detailed analysis of the BSSE results is also reported in this work.

In addition to density functional theory, computations were carried out using Hartree–Fock (HF); Møller–Plesset perturbation theory of order 2 (MP2);⁶⁸ coupled-cluster with single and double excitations (CCSD),⁶⁹ and two methods of coupled-cluster with single, double, and iterative triple excitations CCSD[T] (also known as CCSD+T(CCSD)⁶⁹) and CCSD(T),^{70,71} as implemented in GAMESS. The latter is the highest level of theory applied in this study, and arguably one of the best methods available for single-reference computations.

Several basis sets were employed in this study in order to investigate consistency and predictability across the full set of molecules studied. The basis sets include the correlation consistent basis set of Dunning,⁷² with augmented functions, denoted aug-cc-pVnZ, with $n = D$ for double, T for triple, Q for quadruple, and 5 for quintuple, as implemented in GAMESS (g functions for He and h functions for Ne and Ar are dropped for aug-cc-pV5Z). We note that relative contractions for each split shell differ from He to Ne and Ar and refer readers to the original articles for more details. Extrapolation to the complete basis set (CBS) limit has been carried out for He_3 .

Coupled Cluster Reference Calculations

The potential energy surfaces (PESs) of three Rg_3 systems ($Rg = He, Ne, Ar$) have been investigated with high-level computational methods up to CCSD(T) with complete basis set extrapolation (CBS). The potential energy surface of each D_{3h} Rg_3 molecule is determined with respect to the ground state of the three separated Rg atoms along the radial coordinate of the trimer (Figure 1). In greater detail, the He_3 molecule is used to carefully investigate performance across all methods, the results of which are then extended to the other two systems, Ne and Ar trimers. For highly accurate comparison, we carry out an extrapolation to the complete

basis set limit with coupled cluster, as described in the next sections, for all three trimers.

Convergence Studies for the He₃ Trimer. Accurate calculations have been carried out for He₃ at the CCSD(T) level of theory, establishing convergence of the PES with respect to increasing basis set size, including extrapolation to the complete basis set limit. Additional calculations were carried out with the aug-cc-pVnZ ($n = 2-5$) series to investigate the convergence properties of this family of basis sets. The correlation consistent basis sets of Dunning and co-workers are used to minimize error associated with finite one-particle expansions. These together with extrapolation to the complete basis set (CBS) limit provide high accuracy for electronic energies, enabling quantitative comparison between different *ab initio* methods.

An important component in establishing reliable potential energy surface data involves consistent extrapolation to complete basis set and complete correlation limits.⁷²⁻⁹⁰ While there has been much discussion associated with carrying out complete basis set extrapolation procedures in the literature across a variety of molecular systems, including challenges associated with the type of molecule, family of basis sets being used, treatment of BSSE, and/or properties being extrapolated, the most accurate and reliable extrapolation methodology is not a matter of consensus. In this work, we compare several of the important extrapolation schemes used in the literature, and therefore we first briefly discuss the different approaches in what follows.

The original purpose of Dunning⁷² in the construction of the aug-cc-pVnZ basis sets was to enable the extrapolation of properties using a simple three-points exponential formula, denoted here as $[n,n',n'';Feller]$ -CBS, with $[n,n',n'']$, the cardinality of the employed basis]:

$$f(n) = f^{\text{CBS}} + A \exp(-Bn) \quad (5)$$

where n is the cardinal number of the basis set, for example, $n = 2$ for DZ, 3 for TZ, etc.; $f(n)$ is the property (in this case, energy) obtained using the aug-cc-pVnZ basis set, and f^{CBS} is the extrapolated value for the same property. Several authors assert that eq 5 is suitable for the extrapolation of energies at the Hartree-Fock (HF) level, while in many cases the effective decay for a correlated method (e.g., coupled-cluster) is reasonably slower than the exponential decay.^{75,88}

Many other extrapolation techniques have been developed as alternatives to $[n,n',n'';Feller]$ -CBS. In 1962, Schwartz⁸⁶ proposed an extrapolation procedure for energies of atoms that incorporates an inverse power series function of the basis set extension, n . While this extrapolation has quite a simple expression for atoms, it tends to become very complicated for molecules containing different types of nuclei, requiring further approximations.^{79,91} In the simple case of the helium homotrimer, however, such a formulation can be applied as

$$E(n) = E^{\text{CBS}} + An^{-3} + Bn^{-5} + Cn^{-7} + \dots \quad (6)$$

Truncation of eq 6 leads to simple n -points formulas. In the present study, we denote the simple two-point formula as $[n,n';Schwartz]$ -CBS, the three-point formula as $[n,n',n'';Schwartz]$ -CBS, and so on.

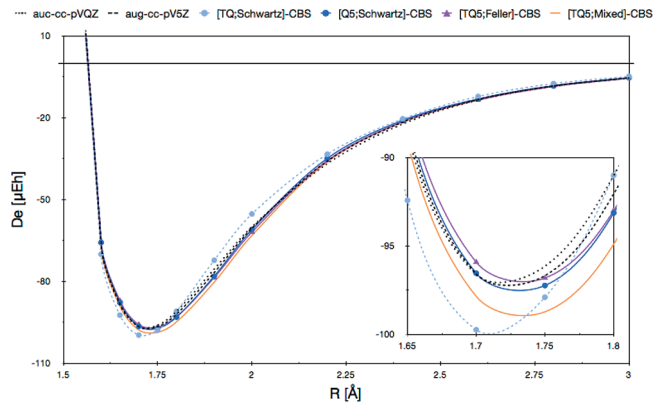


Figure 2. CCSD(T) dissociation energies D_e (μE_h) of the He₃ molecule along the radial coordinate R (Å) as a function of the basis set and CBS extrapolation formula. Dashed black lines are aug-cc-pVQZ and aug-cc-pV5Z.

Table 2. Convergence of Dissociation Energies D_e (μE_h) at the Equilibrium Distances R_{min} (Å), as a Function of Extrapolation Formulas (from above, converging to the middle), and Wave Function Method (from below, converging to the middle)^a

basis set	wave function	D_e (μE_h)	R_{min} (Å)
aug-cc-pVQZ		-96.165	1.72
aug-cc-pV5Z		-96.696	1.72
[TQ;Schwartz]-CBS	CCSD(T)	-99.753	1.71
[Q5;Schwartz]-CBS		-97.253	1.73
[TQ5;Mixed]-CBS		-98.739	1.74
[TQ5;Feller]-CBS	CCSD(T)	-96.829	1.74
	CCSD[T]	-96.843	1.74
[TQ5;Feller]-CBS	CCSD	-84.872	1.74
	MP2	-66.325	1.80
	HF	-0.261	2.80

^a The [TQ5;Feller]-CBS/CCSD(T) method is chosen as the reference for all subsequent calculations.

Also of interest here is the extrapolation technique proposed by Truhlar⁸⁸ and by Halkier et al.,⁷⁵ which couples HF together with correlation methods to obtain a formulation of the type

$$E^{\text{CBS}}(\text{TOT}) = E^{\text{CBS}}(\text{HF}) + E^{\text{CBS}}(\text{corr}) \quad (7)$$

The HF component, $E^{\text{CBS}}(\text{HF})$, and the correlation component, $E^{\text{CBS}}(\text{corr})$, are obtained using different power expansion extrapolations for their respective methods. In the method of Truhlar, a power expansion n^{-A} with variant coefficients is used, while Halkier et al. use an exponential +Schwartz-type expansion. In the present work, we also employ a hybrid three-point Feller exponential formula for the HF component, [TQ5;Feller]-CBS, together with a three-point Schwartz formula, [TQ5-Schwartz]-CBS, for the correlation component. This method will be denoted simply as [TQ5;Mixed]-CBS.

Results for all extrapolation techniques at the CCSD(T) level of theory are presented in the Supporting Information and only summarized here in Figure 2 and Table 2 for the He₃ trimer system. Comparing results without extrapolation using aug-cc-pVQZ and aug-cc-pV5Z basis sets, one observes evidence of reaching the complete basis set limit. A suitable extrapolation technique with these large basis sets

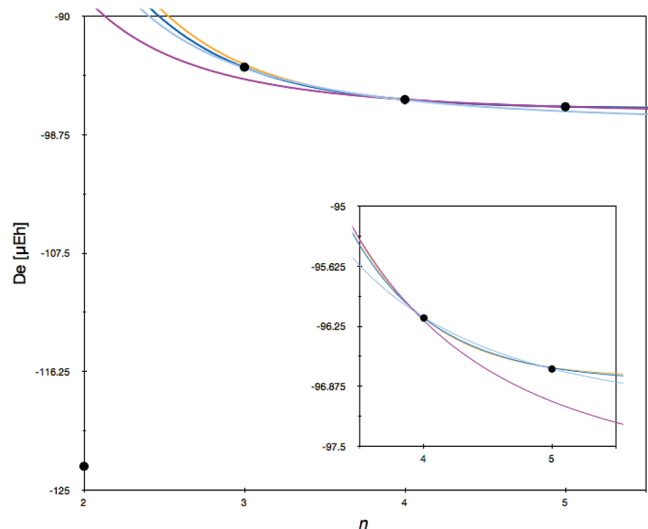


Figure 3. Evaluation of extrapolation formulas at the complete basis set limit as a function of the basis set size, n , for He_3 at $R = 1.75 \text{ \AA}$.

should therefore provide results very close to that of the complete basis set limit. The simple two-point [TQ;Schwartz]-CBS extrapolation method, although an improvement over formulas containing double- ζ basis sets (e.g., see the Supporting Information), which are completely outside a monotonically decreasing behavior, still shows substantial error in the minimum region (Figure 2, inset graph). The best results are achieved with the exponential [TQ5;Feller]-CBS and the simple two-point [Q5;Schwartz]-CBS extrapolation procedures.

Basis set convergence and overall extrapolation behavior can also be viewed by evaluating the interaction energies at a fixed distance near the minimum, $R = 1.75 \text{ \AA}$ (Figure 3). Figure 3 shows that the convergence of the He_3 trimer energy with the aug-cc-pVnZ basis sets is very well approximated by the original exponential formula of Feller. These results are perhaps not a great surprise due to the simple nature of the high-symmetry trimer, with only six electrons. For similar reasons, however, the mixed extrapolation technique appears to be slightly overbound for this simple case, while it might provide more accurate results for more complicated molecular systems. The Feller extrapolation formulation will be used for the other trimer systems in the series.

We next consider optimization of the wave function method, including Hartree–Fock (HF) up to coupled cluster methods. In particular, the dissociation energies, D_e , of the He_3 trimer along the radial coordinate R are calculated using the [TQ5;Feller]-CBS extrapolation with different wave function methods. The full set of computational results across all different wave function types considered can be found in the Supporting Information and is only summarized here in Figure 4 and Table 2.

From the results, we find that HF predicts an essentially unbound system, as expected. MP2 theory shows differences on the order of $30 \mu\text{E}_h$ in the region of the minimum, with respect to the more accurate wave function types. As such, both HF and MP2 wave functions are largely insufficient for the description of the PES of this trimer. A slight improvement is observed with CCSD, which shows a

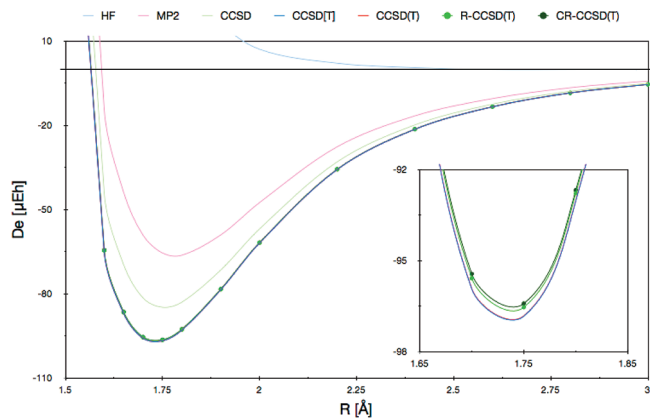


Figure 4. Dissociation energies D_e (μE_h) of the He_3 molecule along the radial coordinate R (Å) using the [TQ5;Feller]-CBS extrapolation procedure as a function of wave function type.

Table 3. CCSD(T) Reference Dissociation Energies D_e (μE_h) of the Rg_3 Molecules along the Radial Coordinate R (Å) Using the [TQ5;Feller]-CBS Extrapolation Procedure

He		Ne		Ar	
R [Å]	D_e [μE_h]	R [Å]	D_e [μE_h]	R [Å]	D_e [μE_h]
1.00	16 907.2	1.00	132 743.7	1.00	1 035 602
1.20	3250.1	1.20	25 043.9	1.20	351 542.3
1.40	400.2	1.40	3794.8	1.40	105 027.2
1.60	-65.200	1.60	57.43	1.60	26 759.2
1.65	-87.120	1.70	-310.907	1.80	4589.3
1.70	-95.933	1.80	-380.858	2.00	-657.443
1.75	-96.829	1.90	-336.865	2.20	-1354.108
1.80	-93.023	2.00	-265.582	2.30	-1247.063
1.90	-78.645	2.20	-155.338	2.40	-1068.008
2.00	-61.991	2.40	-97.182	2.50	-881.626
2.20	-35.794	2.60	-62.580	2.60	-714.031
2.40	-21.478	2.80	-41.109	2.80	-458.702
2.60	-13.472	3.00	-27.786	3.00	-296.944
2.80	-8.537	4.00	-4.331	4.00	-50.669
3.00	-5.508	6.00	-0.342	6.00	-3.599

difference on the order of $12 \mu\text{E}_h$ in the region of the minimum. It is clear that triple excitations are necessary to achieve an accurate description of the trimer system. Both CCSD[T] and CCSD(T) show essentially the same description in the region of the minimum, on the order of $0.02 \mu\text{E}_h$.

CCSD(T)/[TQ5;Feller]-CBS Calculations for He_3 , Ne_3 , and Ar_3 Trimers. From the conclusions obtained for He_3 in the previous sections, the potential energy surfaces for all three Rg_3 's are determined, using the CCSD(T) method and the [TQ5;Feller]-CBS extrapolation, as detailed in Table 3. The predicted minimum radial coordinate distances are 1.75 \AA , 1.80 \AA , and 2.20 \AA , for He_3 , Ne_3 , and Ar_3 , respectively. The dissociation energies show the significant difference in binding characteristics along this series. In particular, while the radial distance difference between the He_3 trimer and Ne_3 trimer is only 0.05 \AA , the difference in dissociation energy is $284 \mu\text{E}_h$. Progressing to the Ar_3 trimer, the radial distance increases by a significant amount, 0.40 \AA (0.45 \AA), as does the dissociation energy, with a difference of $973 \mu\text{E}_h$ ($1257.3 \mu\text{E}_h$), compared to Ne_3 (He_3). These reference calculations will be used in the comparative analysis across density functional classes discussed below.

We should point out that the largest source of error in the reported data in Table 3 is associated with the omission of core/valence correlation energies, which would result in some increase in these numbers. Comparison with the limited data in the literature is very difficult due to the different focus of the studies, resulting in a wide variation in predicted energetics across this set of trimers. This has less to do with the quality of results presented in the literature than the fact that the reference state used is quite different from one investigation to another (e.g., referenced with respect to the energy of the three separated atoms, to the energy of the respective dimer plus atom, to the three-body components, etc.). There is also considerable variance in the reporting of internal coordinates for the systems. This lack of consistency limits reliable comparison, and for our current purpose, it is only necessary to have one consistent set of data to benchmark against the variety of DFT functionals.

One might, however, consider comparison with the corresponding rare gas dimers, which instead emphasizes the much larger energies found for the trimers. Full-CI results of van Mourik and van Lenthe report a binding energy for He₂ of 34.68 μE_h at 2.96 Å,⁹² data in accord with those proposed by Komasa and Rychlewsky using an explicitly correlated Gaussian function approach.⁹³ Other calculated values range from those obtained by Hobza and co-workers (32.2 μE_h with CCSD(T) level, 30.32 μE_h with Full-CI level⁹⁴), to 35.02 μE_h obtained by Szalewicz and co-workers using SAPT calculations.⁹⁵ A more detailed discussion by Specchio et al.⁹⁶ on the full set of available potentials can also be found. An accurate analysis of the potential energy surface of Ne₂ was conducted by Gdanitz.⁹⁷ The reported binding energy is 131.53 μE_h at *R* = 3.10 Å. Other accurate calculations on this system range from the 130.33 μE_h at *R* = 3.10 Å of Cybulski and Toczłowski (aug-cc-pV5Z+bonding function/CCSD(T) level)⁹⁸ to 133.96 μE_h at *R* = 3.09 Å.⁹⁹ Finally, Aziz reported the most accurate value to date for Ar₂, 453.99 μE_h at *R* = 3.75 Å, obtained using a semiempirical potential fit to accurate measured data, within experimental error.¹⁰⁰ With a comparison across this set of reported dimers, one can see a relatively consistent 33–34% increase in interaction energy for the trimers from that of the respective dimer system, which is quite substantial.

DFT Potential Energy Surfaces of Rare Gas Trimers

Performance across several density functional class types for the prediction of the potential energy surfaces of the R_g₃ trimers, as referenced against the accurate CCSD(T) [TQ5;Feller]-CBS extrapolated results, is provided in Table 3. The aug-cc-pVnZ family of basis sets, with *n* = D and T, is used for all reported calculations here, in combination with 34 different exchange-correlation functional approximations. The ultrafine (96, 1202) (corresponds to the “army” grid in GAMESS) and (400, 770) Lebedev grids⁶⁶ have been used for all calculations. Basis set superposition error (BSSE) is accounted for using the counterpoise (CP) method¹⁰¹ and further elaborated upon in the discussion.

To facilitate evaluation of the performance of each functional, a mean absolute deviation with respect to the

accurate CCSD(T)/CBS data (Table 3) has been calculated on the usual 15-point grid used for the PES calculation. For this evaluation, one could argue that the mean absolute deviation, MAD, evaluated as

$$\text{MAD} = \sum_{X=1}^{X_{\text{tot}}} \left| \frac{D_e^{\text{xc}}(X) - D_e^{\text{CCSD(T)}}(X)}{X_{\text{tot}}} \right| \quad (8)$$

is not the best parameter for evaluation of the functional performance, due to the fact that it accounts for equal weighting of each point on the grid. To establish a single reliable evaluation parameter for the problem at hand, errors in the region of the minimum should count more than errors in regions far from the minimum. For example, points in the repulsive region at short-range should be more heavily weighted. For this reason, a weighted mean absolute deviation, denoted wMAD, is calculated in addition to the usual MAD. The wMAD values are calculated by giving a weight factor to each point of the grid according to the distance of that point from the equilibrium, and the relative shape of the accurate PES. The weight factors are calculated with the CCSD(T)/CBS PES by scaling each point according to the relative heights of the PES at that point and by renormalization of the total weight to the total number of points in the grid (see the Supporting Information).

In addition to the mean absolute deviation parameters, the difference from the absolute CCSD(T)/CBS value (deviation from reference, abbreviated DFR) in the region of the minimum (*R* = 1.75 Å for He₃, *R* = 1.80 Å for Ne₃, and *R* = 2.20 Å for Ar₃) is also determined in the evaluation of the performance of each density functional approximation. The MAD and wMAD values with respect to the accurate PES for 34 different DFT functionals and two basis sets are collected in Table 4 for He, Ne, and Ar.

For the He₃ trimer, the two functionals that appear to perform the best are the τ-HCTHhyb and ωB97X functionals. Relative to the majority of the data, reasonable performance is also obtained with several of the other functionals, including B97, B97-2, B98, TPSS, TPSSH, M05 and M05-2X, M08-HX and M08-SO, and ωB97X and ωB97X-D. Interestingly, the semiempirical dispersion correction does not seem to improve the results in the two cases it was used, B97-D and ωB97X-D, compared to their uncorrected counterparts. In several cases, we also note that the BSSE does not necessarily improve the results, a point that will be revisited below.

Moving on to the heavier trimers, Ne₃ and Ar₃, which have larger atomic polarizabilities than He₂, we see a difference in the trends of the set of functionals. This might be anticipated on the basis of the difference in bonding in these heavier trimers. In the case of the Ne₃ trimer, the B97 family of functionals performs very well, particularly the B98, B97, B97-1, and B97-K functionals. Other functionals also have relatively good performance, including M05, M05-2X, M06-HX and M06-SO, TPSS and TPSSH, and τ-HCTHhyb. However, while the DFR value is fairly low, the wMAD values are noticeably large. Additionally, several of these functionals need to be used with caution given the known spurious oscillatory behavior having to do with the kinetic

Table 4. MAD, wMAD, and Deviation from Reference near the Equilibrium Distance for 34 Different Density Functionals and (a) He₃, (b) Ne₃, (c) Ar₃^a

helium trimer	aug-cc-pVTZ		
	MAD	wMAD	DFR <i>R</i> = 1.75
B97	201.0 (201.8)	65.7 (68.9)	-78.2 (-81.7)
B97-1	132.3 (134.3)	113.6 (116.7)	-125.7 (-128.9)
B97-2	453.5 (449.6)	43.3 (39.8)	32.4 (28.7)
B97-3	707.1 (690.5)	518.2 (510.9)	583.7 (576.9)
B97-D	679.6 (681.3)	73.3 (78.6)	-102.4 (-108.9)
B97-K	276.9 (283.6)	118.5 (122.3)	-119.8 (-123.2)
B98	134.5 (135.1)	73.4 (76.5)	-84.2 (-87.5)
HCTH/93	1227.9 (1,220.7)	306.2 (298.8)	318.5 (310.4)
HCTH/120	436.9 (438.1)	318.2 (324.2)	-368.1 (-374.7)
HCTH/147	515.3 (515.5)	149.5 (155.7)	-183.6 (-190.5)
HCTH/407	696.9 (700.3)	637.8 (644.9)	-732.7 (-740.4)
τ -HCTH	487.5 (489.9)	271.0 (277.3)	-325.6(-332.7)
τ -HCTHhyb	313.4 (312.0)	13.9 (15.2)	-17.2 (-21.0)
BMK	1128.9 (1,085.6)	871.6 (854.3)	947.9 (932.3)
ω B97	147.8 (148.2)	185.9 (180.6)	214.1 (208.6)
ω B97X	31.4 (26.9)	13.7 (14.6)	-5.6 (-10.6)
ω B97X-D	549.1 (541.0)	73.3 (71.1)	46.9 (41.8)
BLYP	516.6 (511.1)	400.7 (395.4)	457.1 (451.4)
B3LYP	253.8 (248.9)	258.1 (254.2)	297.7 (293.8)
B2-PLYP	111.4 (112.6)	147.8 (141.6)	174.3 (168.3)
CAMB3LYP	221.7 (221.0)	75.8 (71.2)	95.6 (90.7)
VS98	501.9 (505.4)	224.7 (227.5)	-224.5 (-231.4)
PKZB	796.2 (796.8)	209.3 (216.0)	-256.8 (-264.2)
TPSS	166.7 (166.3)	69.7 (74.4)	-87.1 (-92.6)
TPSSH	141.8 (141.2)	46.0 (49.5)	-58.7 (-63.0)
TPSSM	219.9 (219.3)	87.2 (92.3)	-106.71 (-112.6)
M05	140.3 (148.7)	82.2 (84.5)	-63.1 (-68.8)
M05-2X	128.5 (126.2)	49.6 (49.1)	5.9 (-1.7)
M06	278.9 (250.0)	317.6 (332.6)	-422.9 (-438.9)
M06-2X	226.7 (226.5)	314.8 (319.2)	-383.5 (-387.5)
M06-L	140.7 (109.8)	60.5 (92.9)	-98.7 (-134.5)
M06-HF	497.4 (492.7)	365.7 (382.3)	-422.9 (-439.3)
M08-HX	104.8 (108.7)	65.3 (61.9)	70.1 (62.5)
M08-SO	87.9 (54.8)	37.8 (42.5)	-33.4 (-48.7)

neon trimer	aug-cc-pVTZ		
	MAD	wMAD	DFR <i>R</i> = 1.80
B97	1378.3 (1252.5)	107.2 (51.5)	27.5 (-3.0)
B97-1	808.5 (735.8)	50.7 (101.4)	-12.6 (-41.7)
B97-2	1938.4 (1786.4)	391.3 (323.3)	146.7 (106.9)
B97-3	1786.3 (1652.3)	1082.0 (1024.8)	703.9 (668.8)
B97-D	2850.1 (2726.7)	242.8 (266.5)	-390.4 (-430.1)
B97-K	731.6 (689.0)	147.7 (189.2)	12.0 (-13.4)
B98	980.5 (848.8)	74.3 (21.6)	30.6 (-3.1)
HCTH/93	4,098.9 (3969.8)	990.5 (931.3)	413.6 (381.5)
HCTH/120	1827.5 (1771.6)	316.0 (365.9)	-314.4 (-341.0)
HCTH/147	2328.1 (2231.7)	133.5 (121)	-114.0 (-142.4)
HCTH/407	2417.0 (2374.8)	840.4 (893.3)	-707.6 (-733.5)
τ -HCTH	2157.1 (1987.3)	187.3 (220.8)	-250.1 (-315.0)
τ -HCTHhyb	1404.2 (1233.8)	259.6 (184.0)	90.9 (45.7)
BMK	1439.0 (1242.8)	1831.5 (1750.6)	1033.8 (974.0)
ω B97	503.4 (464.5)	404.6 (358.5)	402.4 (379.2)
ω B97X	418.2 (275.7)	209.4 (147.0)	187.5 (150.8)
ω B97X-D	1580.1 (1372.4)	326.2 (258.9)	-49.3 (-114.8)
BLYP	1855.3 (700.0)	749.6 (700.0)	579.4 (555.7)
B3LYP	926.8 (445.5)	501.6 (445.5)	415.6 (385.9)
B2-PLYP	454.1 (191.7)	287.4 (191.7)	265.2 (200.4)
CAMB3LYP	499.3 (122.2)	137.9 (122.2)	238.7 (211.7)
VS98	1220.3 (444.0)	457.3 (444.0)	573.0 (548.8)
PKZB	3163.1 (137.6)	125.7 (137.6)	-171.4 (-193.5)
TPSS	1383.9 (112.0)	157.3 (112.0)	-11.1 (-52.3)
TPSSH	1182.3 (110.6)	184.0 (110.6)	39.0 (-4.5)
TPSSM	1677.4 (108.4)	138.2 (108.4)	-36.2 (-73.5)
M05	608.8 (193.0)	143.0 (193.0)	79.8 (35.2)
M05-2X	402.3 (219.6)	172.1 (219.6)	134.7 (78.9)

Table 4. Continued

helium trimer	aug-cc-pVTZ		
	MAD	wMAD	DFR <i>R</i> = 1.75
M06	996.4 (269.6)	218.3 (269.6)	-310.1 (-400.5)
M06-2X	552.9 (392.5)	315.5 (392.5)	-306.8 (-356.5)
M06-L	183.7 (303.2)	178.4 (303.2)	-128.4 (-256.4)
M06-HF	551.9 (560.3)	292.9 (560.3)	-410.5 (-655.9)
M08-HX	529.8 (72.5)	214.3 (72.5)	87.4 (-21.3)
M08-SO	590.8 (47.1)	108.4 (47.1)	15.6 (-52.4)

argon trimer	aug-cc-pVTZ		
	MAD	wMAD	DFR <i>R</i> = 2.20
B97	2912.7 (2837.5)	916.3 (848.6)	1268.1 (1186.7)
B97-1	2028.4 (1952.3)	513.2 (445.3)	643.2 (561.0)
B97-2	4035.9 (3955.6)	1466.5 (1395.3)	2101.4 (2015.9)
B97-3	3701.8 (3486.0)	1981.4 (1889.8)	2573.0 (2458.1)
B97-D	4722.7 (4670.1)	377.2 (350.0)	528.0 (439.7)
B97-K	5986.0 (5850.1)	460.5 (403.2)	541.4 (470.4)
B98	2286.4 (2200.8)	773.4 (698.6)	1041.9 (951.7)
HCTH/93	7934.2 (7868.7)	2690.4 (2626.7)	3905.5 (3830.6)
HCTH/120	4941.3 (4938.9)	363.0 (334.0)	539.2 (485.2)
HCTH/147	5610.7 (5,559.5)	1008.0 (956.0)	1479.6 (1418.9)
HCTH/407	5901.9 (5900.4)	733.8 (767.0)	-614.0 (-665.6)
τ -HCTH	6003.0 (5985.4)	773.9 (675.1)	1204.6 (1089.7)
τ -HCTHhyb	3191.2 (3169.4)	1146.1 (1056.9)	1617.8 (1511.5)
BMK	5738.6 (5302.1)	3258.8 (3090.5)	4430.9 (4217.3)
ω B97	1015.7 (1021.3)	658.0 (616.8)	581.1 (489.8)
ω B97X	729.9 (733.2)	242.6 (205.4)	80.9 (-9.9)
ω B97X-D	2114.3 (1991.7)	666.3 (569.4)	1087.3 (957.5)
BLYP	4259.9 (4,185.0)	2293.3 (2226.4)	3077 (2998.8)
B3LYP	3402.2 (3248.0)	1780.2 (1713.6)	2369.9 (2290.3)
B2-PLYP	2517.8 (1979.8)	1086.3 (970.6)	1426.7 (1278.9)
CAMB3LYP	1477.1 (1331.0)	1110.6 (1049.7)	1418.5 (1345.6)
VS98	2732.7 (2706.4)	1778.7 (1813.2)	-2847.6 (-2905.2)
PKZB	5547.3 (5502.8)	1089.1 (1041.2)	1615.0 (1559.5)
TPSS	3294.5 (3268.6)	1204.5 (1119.3)	1766.3 (1665.3)
TPSSH	2822.0 (2794.2)	1208.2 (1123.2)	1751.3 (1650.8)
TPSSM	3455.6 (3371.8)	1194.2 (1116.3)	1753.4 (1661.5)
M05	3157.4 (3072.1)	443.6 (364.9)	425.7 (334.0)
M05-2X	1054.6 (753.1)	196.6 (120.3)	87.3 (-62.6)
M06	1724.4 (1497.2)	610.6 (439.3)	897.1 (646.2)
M06-2X	621.8 (503.3)	241.0 (167.3)	472.5 (346.0)
M06-L	1456.6 (1157.1)	347.0 (113.5)	436.9 (42.5)
M06-HF	1299.9 (1678.5)	534.8 (351.2)	878.4 (511.3)
M08-HX	1645.6 (1337.4)	719.6 (570.5)	797.6 (598.2)
M08-SO	1639.2 (1279.9)	463.6 (284.3)	496.0 (272.0)

^a Results shown are BSSE corrected (BSSE uncorrected).

energy density component.^{102–104} This is addressed in more detail in the next section.

Finally, for the Ar₃ trimer, we notice that considerably fewer functionals provide reasonable performance. In this case, the ω B97X functional stands out, with the M05-2X functional being also reasonable relative to the other functionals. This would indicate a type of bonding in this trimer that is not well represented by most of these functionals.

Across all trimers, several meta-GGA functionals, such as VS98 and PKZB, show results that are highly dependent on the system as well as the basis set, with acceptable results in many cases, but quite poor results in others. As with the other meta-GGA functionals, these two functionals also show oscillating behavior. The BLYP, B3LYP, CAM-B3LYP, and BMK functionals also have overall poor performance, something that could be a result of specialized parametrization for specific properties, e.g., for kinetic data, rather than

Table 5. M06-L/aug-cc-pVTZ Energy, RMS Gradient, and Computational Time for Ar₃ Trimer, as a Function of Grid Specification

grid	grid specification		$R = 1.00 \text{ \AA}$			$R = 2.2 \text{ \AA}$		
	radial	angular	energy	RMS	CPU	E	RMS	CPU
Lebedev SG1	24	1–94 (variable)	–1581.58670477	0.5132730	17.9		D.N.C. ^a	
polar coordinate	96	Theta = 12, Phi = 24	–1581.58369967	0.5142771	11.1	–1582.61268820	0.0004244	5.1
Lebedev default	96	302	–1581.58406747	0.5136053	12.1		D.N.C.	
Lebedev R1,tight	96	590	–1581.58406691	0.5138834	16.7	–1582.61268238	0.0002371	24.6
Lebedev R1,U-tight	96	770	–1581.58409109	0.5139260	17.7	–1582.61270112	0.0002315	31.5
Lebedev R2,U-tight	200	770	–1581.58409833	0.5139036	42.7	–1582.61259280	0.0001795	61.5
Lebedev R3,U-tight	250	770	–1581.58409854	0.5139029	32.6	–1582.61259277	0.0001564	41.3
Lebedev R4,U-tight	300	770	–1581.58409812	0.5139005	46.3	–1582.61259271	0.0001701	71.7
Lebedev R5,U-tight	400	770	–1581.58409862	0.5138996	63.3	–1582.61259271	0.0001701	71.7
Lebedev Army	96	1202	–1581.58409655	0.5139938	31.9	–1582.61269837	0.0002469	39.6
Lebedev R6, Army	250	1202	–1581.58410481	0.5139705	46.4	–1582.61259482	0.0001570	64.1

^a SCF procedure did not converge.

structural, in the case of BMK. The B2-PLYP functional also provides relatively unsatisfactory results in all cases.

Quality of Density Functional Theory Integration Grid. The noted oscillatory behavior observed in several regions of the PES for some of the represented density functionals, in particular the meta-GGA functionals, is not an unknown phenomenon within the context of density functional theory and has been discussed in the literature.^{102,103,105–107} Such behavior was discussed in a recent article by Johnson and co-workers,¹⁰² specifically in reference to the meta-GGA functional analysis of the PES for a set of dispersion-bound complexes, including the Ar trimer. This erroneous behavior is thought to originate from the divergence of the τ -dependent term in these functionals. The suggestion from Johnson and co-workers of how to avoid this spurious behavior is to use an ultrafine grid.

To illustrate the effect of grid specification, we have carried out computations using a variety of standard as well as other more and less stringent grid specifications, specifically for the Ar₃ system and the M06-L functional, which showed oscillatory behavior even when using the relatively stringent army grade grid (96, 1202). Table 5 shows the results for two geometry specifications of the Ar₃, one relatively far from the PES minimum, $R = 1.0 \text{ \AA}$, and a geometry near the PES minimum, $R = 2.2 \text{ \AA}$, where we find a region of some functional oscillation problems. Two types of grid quadratures are represented in the table, an older polar coordinate grid¹⁰⁸ and the newer Lebedev grid.⁶⁶ The latter grid has been found to be more efficient for use in DFT, due to the reduction in the number of quadrature points needed to obtain convergence, compared to other grids.^{108–110} The exchange, correlation, and kinetic energy correction contributions to the energy are determined by summing the contributions from grids centered on each nuclei.^{111,112} The quadrature specification for the angular component is combined with the 1D integration for the radial component and enables the numerical construction of the required integrals. The polar grid specification has been predominantly replaced by the Lebedev grid due to the rather poor distribution of points on the spherical grids, which requires considerably more grid points to obtain a reasonable invariance to rotation.

Analysis of the data in Table 5 shows the sensitivity of the grid to energy, gradient, and associated computational cost, for this functional. In the region specified by $R = 1.00$

\AA , a reasonable choice of angular and radial specification (e.g., at least the default (96, 302) specification), shows quite good convergence. Use of a very loose grid clearly results in poor representation of the PES. However, near the observed oscillatory behavior in the PES, $R = 2.2 \text{ \AA}$, one readily sees the need for a tighter grid specification, particularly associated with the angular component. An optimal tight grid choice is typically observed with an angular component of at least 770. Therefore, using 770 or the typical “army grid” specification of 1202 is expected to provide relatively good convergence in structure and property.

The radial grid and associated weights are a function of the Bragg–Slater radius of the atom, and therefore the number of grid points is expected to vary with the atom type. While typical values are 96 (or 99) for most molecules investigated, higher specifications (e.g., 150–250) may be necessary with heavier atoms or, likewise, much smaller grid specifications for very small atoms. However, if overly large radial specifications are made for the particular atom, the points can become so dense and the extent of the spheres so extreme that numerical instabilities can be observed. This is the case, for example, in the raw data of Table 5, where one sees small oscillations in the data beyond a radial specification of about 250. The effect appears to be much smaller than for the angular component (e.g., Figure 5).

Results displayed in Figure 5 show a comparison of data from (400, 770) and (96, 1202) grid specifications for the M06-L functional. Despite the fact that this more extreme nonstandard radial specification provided by a (400, 770) grid provides a much smoother shape of the PES (and in general, for functionals that have oscillations in some areas of the PES), the underlying behavior and overall performance of the affected functionals does not improve. For example, wMAD for He₃ using M06-L is 60.5 with an army grade grid (96, 1202) and 62.3 with a (400, 770) grid; for the TPSS functional, wMAD is 166.7 and 166.1, respectively. This could be indicative of an overly large value for the radial component, which does result in a smoothing of the PES with the use of the tighter grid but likely has too many points, resulting in a small degradation in overall performance.

In a more recent publication by Wheeler and Houk¹⁰⁴ focusing only the M06 suite of functionals, it was concluded that such spurious behavior originates from grid sensitivity in the kinetic energy density enhancement factor used in the

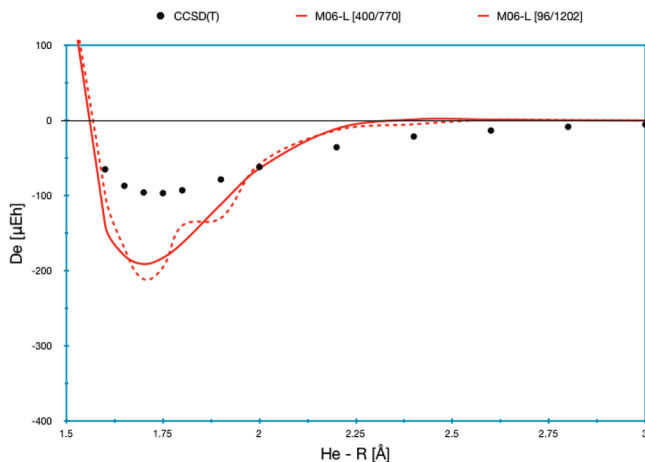


Figure 5. Effect grid size with the M06-L functional for the He_3 potential energy surface.

exchange component of these functionals. While it is clear that the grid specification is quite important, the balance of radial and angular grid specification probably warrants a more systematic investigation, particularly as it is an inherent problem of the meta-GGA class of functionals. In general, sufficiently tight grids should always be used when calculating structure and properties to avoid erroneous behavior.

Assessment of DFT Surfaces. For a global view of the behavior of the series of density functionals, Figure 6 shows the results of PES data for the set of density functional class types. The first class consists of nonhybrid functionals, including B97-D, HCTH/407, and BLYP, as well as τ -HCTH, TPSSH, and M06-L, that are nonhybrid meta-GGAs. The second class consists of a selection of hybrid functionals, as well as range-separated hybrids, including the most complete reparameterization of the B97 functional (called B98), the commonly used B3LYP functional, the double-hybrid B2-PLYP, the range-separated CAM-B3LYP, and the new ω B97 functionals of Chai and Head-Gordon.³² ω B97, ω B97X, and ω B97X-D. The third class includes a selection of hybrid meta-GGA functionals, including TPSSH, M05, M05-2X, M06, M06-2X, M08-HX, and M08-SO.

For helium, very few functionals provide a dissociation curve close to the reference data. The ω B97X functional is the only functional that comes close to the reference data. Within the local functionals, the meta-GGA M06-L (only with a tight radial specification) and TPSS are the only functionals of the set that present at least a reasonable overall curve shape, albeit strongly overbinding. The B97-D curve is shifted to longer equilibrium geometry and also is strongly overbinding. Many of the common GGA functionals are either dissociative or weakly bound for this system. It is interesting to compare ω B97X, which does a good job close to the minimum, with the dispersion corrected version, ω B97X-D, which shows a drastic shift to longer equilibrium but now has the correct dissociation. The class of hybrid meta-GGA functionals has quite widely variant behavior from one functional to another, making it unclear if any particular result is fortuitous or due to a correct description of the physics. All essentially have defects that would be unacceptable for accurate prediction, which is perhaps disappointing considering the formulation of these function-

als. The TPSSH functional has the most reasonable prediction of the whole set, but it is strongly overbinding.

Looking at the classes of functionals for Ne_3 , we observe some improvements in functional performance, which may be attributed to the increased binding energy for this heavier rare gas trimer. The M06-L (only with a tight radial grid) and TPSS nonhybrid meta-GGA's again show the most reasonable overall binding curves, albeit now the TPSS functional is slightly displaced to a longer equilibrium position and is underbinding. The B98 hybrid GGA functional is also reasonable, yet slightly underbound. The ω B97X functional in this system now is considerably underbinding, which is corrected with the semiempirical dispersion, but again ω B97X is shifted to a longer equilibrium distance. The performance of meta-GGA functionals is once again far from acceptable, despite some small general improvements.

Finally, we look at the three classes of functionals for Ar_3 , which has the largest atomic polarizability of the series. This fact is reflected in the behavior of many functionals, which show more reasonable binding curves in comparison to the other two trimers. In particular, B97-D, M06-L (with a tight radial grid), and τ -HCTH nonhybrid functionals perform well, but HCTH/407 is still overbinding, and TPSS is progressively more under-binding than in Ne_3 . Of the hybrid GGA functionals, ω B97X and ω B97-D show again this trend of the former having reasonable prediction around the equilibrium geometry, and the latter only having reasonable dissociation but considerably underbinding. Mostly all of the meta-GGA functionals are underbinding, but many more have overall correct behavior, predicting curves that are at least within the region of the correct reference data, indicating a more correct description of the physics of this system.

Across all systems, the ω B97X functional (with the exception of an estimated underbinding curve for Ne_3) and the M06-L nonhybrid meta-GGA functional with a sufficiently tight radial grid, show a relatively consistent performance across all trimers. It is evident that the choice of the functional as well as the grid extent must be seriously considered for reliable results. It is interesting to note the general very poor behavior of some commonly used functionals, for example, BLYP, B3LYP, and the double hybrid B2-PLYP. For all three rare gas trimers, these three functionals predict a dissociative or very underbinding phenomenon, indicating more general problems than associated with the variation in binding phenomena in the three systems.

Elaboration on BSSE. Looking carefully at the results in Table 4, in some cases, the BSSE uncorrected results are moderately better than those that are BSSE corrected. However, it is well-known that BSSE plays a key role for accurate treatment of weakly bound complexes, where too small basis sets result in poor prediction of binding energies and intermolecular distances.^{113,114} Although it is generally believed that DFT is much less affected by BSSE than other wave function types, a basis set of at least triple- ζ quality is typically necessary to significantly reduce the BSSE.¹¹⁵ The counterpoise correction (CP)⁶⁷ is a standard method used to correct for BSSE, and while the procedure itself has an associated error (typically results in an overestimation of

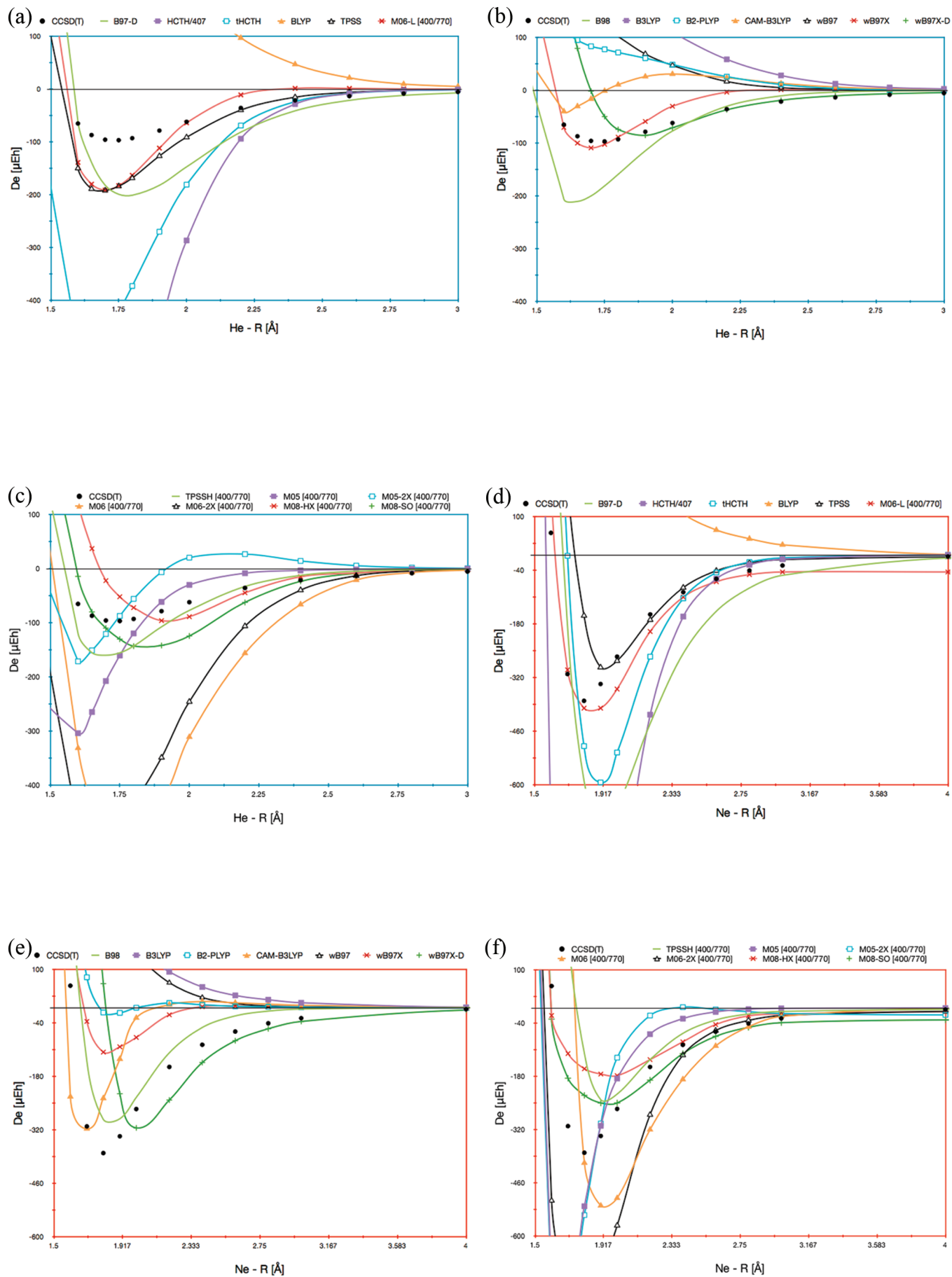


Figure 6

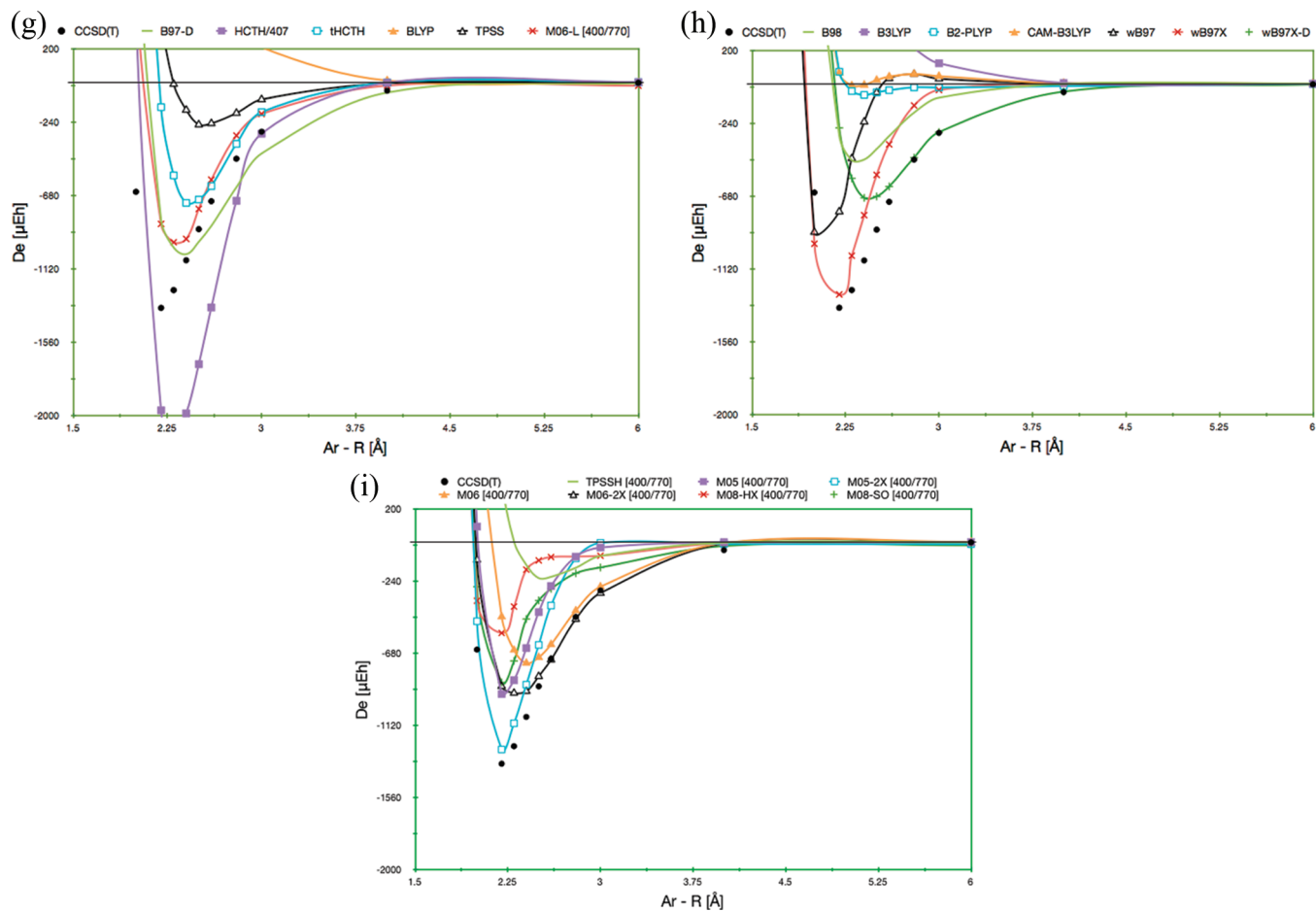


Figure 6. Dissociation energies, D_e (μE_h) of the Rg_3 trimers [a–c, He (blue); d–f, Ne (red); and g–i, Ar (green)] along the radial coordinate R (\AA) using the aug-cc-pVTZ basis set together with local GGA [a, d, g], hybrid GGA [b, e, h], and hybrid meta-GGA [c, f, i] DFT functionals, referenced against the CCSD(T)/CBS results in Table 3. Results are presented on the same relative scale; the (96, 1202) Lebedev grid has been used for all but the meta-GGA class of functionals, for which the (400,770) Lebedev grid is used due to the noted oscillatory behavior in these functionals.

BSSE) and requires additional effort, the method generally provides good results for the vast majority of cases where it is used.

Table 6 shows the BSSE as a function of interatomic distance across five common DFT functionals using the aug-cc-pVDZ basis set for the He_3 trimer. The data are expressed as both an absolute value and a percentage of the total binding energy. The absolute value of the BSSE at the binding distance is, in general, larger than $13 \mu E_h$ for all of the considered functionals, in the order $M06 > B2-PLYP > BMK > BLYP > B3LYP > 13 \mu E_h$. Considering that the dispersion energy near the equilibrium distance, evaluated using the semiempirical dispersion formula of Grimme³³ (namely, the unscaled $-D$ contribution of the B97-D functional), is $-140.2 \mu E_h$, the BSSE energy is always larger than 10% of the dispersion energy. With the exception of the B2-PLYP method, the value of the BSSE in the region close to the minimum represents about 5–7% of the total binding energy. The B2-PLYP method, on the other hand, shows a value considerably larger, representing about 35% of the binding energy, with a large basis set dependency. This is most likely due to the presence of the MP2 term. The general behavior of the BSSE with respect to the interatomic distance is fairly unpredictable, ranging signifi-

cantly across the various functionals, as perhaps seen more clearly in Figure 7. These results show that BSSE plays a substantial role in the determination of the binding energies of the rare gas trimers, particularly with double- ζ quality basis sets, although to a different extent across the functionals.

A reasonable question on the noted effective performance of the BSSE uncorrected combination of DFT and double- ζ quality basis sets arises from this analysis. The proliferation of the use of DFT/double- ζ methods in computational chemistry is primarily due to the relatively cheap computational cost of the combination for a general good performance. However, in many cases, the good performance is amplified by a cancellation of errors in the parametrization of the DFT functional and any associated semiempirical dispersion correction, in combination with the BSSE and the incompleteness of the basis set. In particular, we have previously shown, for small basis sets, that BSSE can be on the same order of magnitude as the dispersion corrections in those functionals that have semiempirical corrections, but the two corrections have different asymptotic behaviors.⁴¹ As such, care should be taken in the selection of the method and basis set, in particular, for computations involving weak interactions. No additional information is revealed in a similar analysis of BSSE effects in the description of the PESs of the other two trimers, Ne_3 and Ar_3 . However,

Table 6. BSSE as a Function of Interatomic Distance for Five Common DFT Functionals, Expressed As Absolute Value in μE_h , ABS, and as Percentage of the Total Binding Energy, %

<i>R</i>	BLYP		B3LYP		BMK		M06		B2-PLYP	
	ABS	%	ABS	%	ABS	%	ABS	%	ABS	%
1	60.24	0.31	63.81	0.36	161.38	0.85	295.65	1.80	179.00	1.05
1.2	29.84	0.64	22.38	0.59	30.92	0.52	71.28	2.15	68.56	2.05
1.4	34.63	2.47	29.01	3.15	26.14	0.80	48.94	27.40	54.96	9.08
1.6	24.75	4.30	22.16	6.85	26.98	1.57	54.80	13.84	42.92	30.40
1.65	21.89	4.60	19.16	7.26	25.59	1.88	52.62	11.07	38.14	35.25
1.7	19.26	4.83	15.99	7.27	22.98	2.22	47.98	8.91	32.83	37.52
1.75	17.08	5.00	13.01	6.88	19.46	2.61	41.55	7.38	27.43	36.23
1.8	15.52	5.36	10.53	6.54	15.54	2.92	34.23	7.45	22.40	34.50
1.9	14.34	6.79	7.64	6.43	8.34	3.43	20.31	5.12	14.64	29.74
2	14.92	9.73	7.24	8.37	3.74	4.07	10.49	3.77	10.35	28.80
2.2	15.69	19.45	8.63	19.36	1.37	30.43	2.85	2.21	7.76	45.87
2.4	12.58	28.70	7.74	33.34	1.49	39.72	1.37	2.34	6.18	84.21
2.6	8.08	32.54	5.28	41.92	1.17	38.03	0.68	2.81	4.09	130.13
2.8	4.50	30.52	3.09	42.36	0.75	41.23	0.33	2.95	2.37	165.88
3	2.24	23.91	1.61	35.05	0.44	53.45	0.19	3.48	1.25	159.86

all tables and graphics associated with this analysis are available as Supporting Information, for those interested in the details.

Conclusions

A systematic investigation into potential energy surfaces characteristics of a series of rare gas trimers across a wide range of methodologies has been presented. Because of the much smaller amount of literature compared to that for rare gas dimer counterparts, it is of interest to investigate this series for a better understanding of *n*-body interaction energies, but also for the evaluation of model chemistries.

In the equilateral D_{3h} case, the trimer series presents a simple one-dimensional potential energy surface, from which the methodology can be compared. These systems represent a challenging test for new methodologies, as all three trimers are van der Waals, or dispersion-bound, systems, with varying degrees of atomic polarizability.

For this study, we have implemented a large set of Kohn–Sham DFT density functionals into the GAMESS software package to fully test performance across a wide range of functional class types, including several of the new dispersion enabled functional strategies. In the process, we also facilitate future implementations and parameter testing of a variety of density functional types. Reference data for the He_3 trimer are investigated in detail using CCSD(T) dissociation energies, D_e (μE_h), of the trimer along the radial coordinate R (\AA) for the aug-cc-pVnZ ($n = 2-5$) series of basis sets and CBS extrapolation. Several well-established extrapolation procedures are compared. Optimal results are achieved with the exponential [TQ5;Feller]-CBS and the simple two-point [Q5;Schwartz]-CBS extrapolation procedures. The [TQ5;Feller]-CBS was subsequently used to establish reference calculations for all three rare gas trimer systems.

Benchmarked against the reference data, investigation is then made across a set of 34 DFT functionals of varying classes, evaluated on the same PES points, using double- and triple- ζ quality basis sets, for all three rare gas trimers. Results with and without correction for basis set superposition error are discussed. Because of spurious oscillation in the potential energy surfaces obtained with meta-GGA functionals, a detailed investigation of the DFT integration grid was also carried out. The tightest grid, (400, 700), was used for all meta-GGA reported results reported here. In general, we propose a sequence of increasing accuracy in terms of (radial, angular) points for Lebedev-type integration grids, as (96,302), (125,590), (250,770), and potentially (400,770), however such a large radial component can easily be overkill, resulting in no additional convergence, but considerable CPU time.

Criteria for the evaluation of calculated potentials for any method on systems of this type should encompass good prediction of the equilibrium distance, dissociation energy,

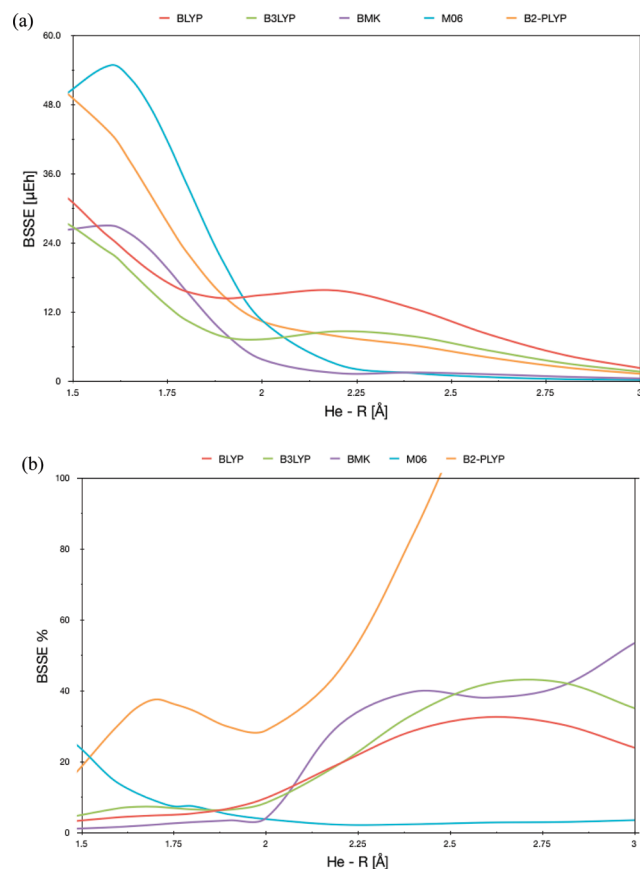


Figure 7. BSSE as a functional of interatomic distance, expressed as (a) absolute value and (b) percentage of the total binding energy.

and long-range behavior. However, what we see from the reported results across DFT classes is that predictions strongly vary depending on the rare gas system, such that no unique best functional can be recommended across the whole series. Local functionals provide a generally poor behavior across all criteria, while hybrid GGAs are somewhat better overall. Meta-GGA functionals are largely unsatisfactory but still represent the best functionals in specific cases, for one or more of the criteria. Oscillations in the behavior of the PES are found for many meta-GGA functionals, despite the use of an ultrafine Lebedev integration grid. We therefore suggest that these functionals be used with caution. A correct long-range behavior should fit to a $c_6/r^6 + c_8/r^8$ like behavior, as is expected from the semiempirical, -D, corrected versions of the functionals, which do quite well for the Rg₂ dimer systems. However, only a few of the potentials calculated with the semiempirically corrected functionals have better long-range behavior, probably the best example being ω B97X-D. The newly implemented range-separated hybrids and semiempirical corrected functionals project the most reasonable global results, relatively speaking. The same does not hold for the B2-PLYP double hybrid functional, which performs quite poorly for all considered cases. The trends in BSSE were considered in more detail, posing a reasonable question on the effective performance of the BSSE uncorrected combination of DFT with double- ζ quality basis sets.

Acknowledgment. We gratefully acknowledge the University of Zürich and the Swiss National Science Foundation for support of this research. We thank M. W. Schmidt for his helpful discussions.

Supporting Information Available: Summary of newly implemented DFT functionals in GAMESS, categorized by DFT class type; CCSD(T) dissociation energies, D_e (μE_h), of the He₃ molecule along the radial coordinate R (Å) for various basis sets and CBS extrapolation formulas; weights used in the wMAD in accord with the shape of the accurate CCSD(T)/CBS PES; MAD, wMAD, and deviation from reference near the equilibrium distance for 34 different density functionals; BSSE as a function of interatomic distance for five common DFT functionals. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Efimov, V. *Phys. Lett.* **1970**, *33B*, 563.
- (2) Efimov, V. *Nucl. Phys. A.* **1973**, *210*, 157.
- (3) Efimov, V. *Comments Nucl. Part. Phys.* **1990**, *19*, 271.
- (4) Suno, H.; Esry, B. D. *Phys. Rev. A* **2008**, *78*, 062701.
- (5) Richard, J. M. *Few-Body Syst.* **2006**, *38*, 79–84.
- (6) Zhukov, M. V.; Danilin, B. V.; Fedorov, D. V.; Bang, J. M.; Thompson, I. S.; Vaagen, J. S. *Phys. Rep.* **1993**, *231*, 151.
- (7) Cybulski, S. M.; Toczylowski, R. R. *J. Chem. Phys.* **1999**, *111*, 10520.
- (8) Bressanini, D.; Morosi, G. *Phys. Rev. A* **2003**, *90*, 133401.
- (9) Bressanini, D.; Morosi, G. *Few-Body Syst.* **2004**, *34*, 1–3.
- (10) Cencek, W.; Jeziorska, M.; Akin-Ojo, O.; Szalewicz, K. *J. Phys. Chem. A.* **2007**, *111*, 11311.
- (11) Cencek, W.; Patkowski, K.; Szalewicz, K. *J. Chem. Phys.* **2009**, *131*, 064105.
- (12) Giese, T.; York, D. M. *Int. J. Quantum Chem.* **2004**, *98*, 388–408.
- (13) Giese, T.; York, D. M. *J. Chem. Phys.* **2004**, *120*, 590.
- (14) Ichihara, A.; Itoh, A. B. *Chem. Soc. Jpn.* **1990**, *63*, 958–960.
- (15) Kim, Y. S. *Phys. Rev. A* **1975**, *11*, 796–803.
- (16) Blume, D.; Greene, C. H.; Esry, B. D. *J. Chem. Phys.* **2000**, *113*, 2145.
- (17) Chakravarty, C.; Hinde, R. J.; Leitner, D. M.; Wales, D. J. *Phys. Rev. E* **1997**, *56*, 563.
- (18) González-Lezana, T.; Rubayo-Soneira, J.; Miret-Artés, S.; Gianturco, F. A.; Delgado-Barrio, G.; Villarreal, P. *J. Chem. Phys.* **1999**, *110*, 9000.
- (19) Bruch, L. W.; Novaro, O.; Flores, A. *J. Chem. Phys.* **1977**, *67*, 2371.
- (20) Lim, T. K.; Duffy, K.; Nakaichi, S.; Akaishi, Y.; Tanaka, H. *J. Chem. Phys.* **1979**, *70*, 4782.
- (21) Sandhas, W.; Kolganova, E. A.; Ho, Y. K.; Motovilov, A. K. *Few-Body Syst.* **2004**, *34*, 137.
- (22) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103/1–161103/4.
- (23) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.
- (24) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (25) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849.
- (26) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559–9569.
- (27) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.
- (28) Leininger, T.; Stoll, H.; Werner, H. J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151–160.
- (29) Yanai, T. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (30) Hetzer, G.; Pulay, P.; Werner, H.-J. *Chem. Phys. Lett.* **1998**, *290*, 143.
- (31) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 7274.
- (32) Chai, J. D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *108*, 084106.
- (33) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (34) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108–034115.
- (35) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398–4401.
- (36) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A.* **2004**, *108*, 4786.
- (37) Tarnopolsky, A.; Karton, A.; Sertchook, R.; Vuzman, D.; Martin, J. M. L. *Phys. Chem. A* **2008**, *112*, 3–8.
- (38) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401.
- (39) Bode, B. M.; Gordon, M. S. *Mol. Graphics Modell.* **1999**, *16*, 133–138.

- (40) Tarini, M.; Cignoni, P.; Montani, C. *IEEE Trans. Visual. Comput. Graphics* **2006**, *12*, 1237–1244.
- (41) Peverati, R.; Baldridge, K. K. *J. Chem. Theory Comput.* **2008**, *4*, 2030–2048.
- (42) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554–8560.
- (43) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264–6271.
- (44) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.
- (45) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233–9242.
- (46) Strange, R.; Manby, F. R.; Knowles, P. J. *Comput. Phys. Commun.* **2001**, *136*, 310–318.
- (47) Miehlich, B.; Stoll, H.; Savin, A. *Mol. Phys.* **1997**, *91*, 527.
- (48) Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624–9631.
- (49) Boese, A. D.; Doltsinis, N. L.; Handy, N. C.; Sprik, M. *J. Chem. Phys.* **2000**, *112*, 1670–1678.
- (50) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2001**, *114*, 5497–5503.
- (51) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (52) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (53) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200–206.
- (54) Stephens, P. J.; Devlin, F. J.; Chablowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (55) Hertwig, R. H.; Koch, W. *Chem. Phys. Lett.* **1997**, *268*, 345–351.
- (56) van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400–410.
- (57) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. *Phys. Rev. Lett.* **1999**, *82*, 2544–2547.
- (58) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401–146404.
- (59) Perdew, J. P.; Tao, J. M.; Staroverov, V. N.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 6898–6911.
- (60) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129–12137.
- (61) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2004**, *121*, 11507.
- (62) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Csonka, G. I.; Scuseria, G. E. *Phys. Rev. A* **2007**, *76*, 042506–042511.
- (63) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 1009–1018.
- (64) Zhao, L.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101–194119.
- (65) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126–13130.
- (66) Lebedev, V. I.; Laikov, D. N. *Dokl. Math.* **1999**, 477–478.
- (67) Su, P.; Li, H. *J. Chem. Phys.* **2009**, *131*, 014102.
- (68) Möller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
- (69) Piecuch, P.; Kucharski, S. A.; Kowalski, K.; Musial, M. *Comput. Phys. Commun.* **2002**, *149*, 71–96.
- (70) Bentz, J. L.; Olson, R. M.; Gordon, M. S.; Schmidt, M. W.; Kendall, R. A. *Comput. Phys. Commun.* **2007**, *176*, 589–600.
- (71) Olson, R. M.; Bentz, J. L.; Kendall, R. A.; Schmidt, M. W.; Gordon, M. S. *J. Comput. Theor. Chem.* **2007**, *3*, 1312–1328.
- (72) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (73) de Lara-Castellis, M. P.; Krems, R. V.; Buchachenko, A. A.; Delgado-Barrio, G.; Villarreal, P. *J. Chem. Phys.* **2001**, *115*, 10438.
- (74) Feller, D. F.; Sordo, J. A. *J. Chem. Phys.* **2000**, *112*, 5604.
- (75) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsena, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243.
- (76) Martin, J. M. L. *Chem. Phys. Lett.* **1996**, *262*, 97–104.
- (77) Martin, J. M. L. *Chem. Phys. Lett.* **1996**, *259*, 669–678.
- (78) Martin, J. M. L. *Chem. Phys. Lett.* **1998**, *292*, 411.
- (79) Martin, J. M. L.; Taylor, P. R. *J. Chem. Phys.* **1997**, *106*, 8620–8623.
- (80) Nyden, M. R.; Petersson, G. A. *J. Chem. Phys.* **1981**, *75*, 1843.
- (81) Peterson, K. A.; Dunning, T. H., Jr. *J. Phys. Chem. A* **1997**, *101*, 6280–6292.
- (82) Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 7410.
- (83) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A.; Mantzaris, J. *J. Chem. Phys.* **1988**, *89*, 2193.
- (84) Petersson, G. A.; Frisch, M. J. *J. Phys. Chem. A* **2000**, *104*, 2183.
- (85) Petersson, G. A.; Yee, A. K.; Bennett, A. *J. Chem. Phys.* **1985**, *83*, 5105.
- (86) Schwartz, C. *Phys. Rev.* **1962**, *126*, 1015.
- (87) Schwartz, C. *Computational Physics*; Academic: New York, 1963; Vol. 2.
- (88) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *294*, 45–48.
- (89) Wilson, A. K.; Dunning, T. H., Jr. *J. Chem. Phys.* **1997**, *106*, 8718.
- (90) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 2975.
- (91) Martin, J. M. L. *Chem. Phys. Lett.* **1996**, *259*, 679–682.
- (92) van Mourik, T.; van Lenthe, J. H. *J. Chem. Phys.* **1995**, *102*, 7479.
- (93) Komasa, J.; Rychlewski, J. *Mol. Phys.* **1997**, *91*, 909.
- (94) Burda, J. V.; Zahradnik, R.; Hobza, P.; Urban, M. *Mol. Phys.* **1996**, *89*, 425.
- (95) Korona, T.; Williams, H. L.; Bukowski, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **1997**, *106*, 5109.
- (96) Specchio, R.; Famulari, A.; Raimondi, M. *THEOCHEM* **2001**, *549*, 77.
- (97) Gdanitz, R. *J. Chem. Phys. Lett.* **2001**, *348*, 67.
- (98) Cybulski, S. M.; Toczłowski, R. R. *J. Chem. Phys.* **1999**, *111*, 10520.
- (99) Aziz, R. A.; Slaman, M. *Chem. Phys.* **1989**, *130*, 187–194.
- (100) Aziz, R. A. *J. Chem. Phys.* **1993**, *99*, 4518.
- (101) Jansen, H. B.; Ross, P. *Chem. Phys. Lett.* **1969**, *3*, 140.
- (102) Johnson, E. R.; Becke, A. D.; Sherrill, C. D.; DiLabio, G. A. *J. Chem. Phys.* **2009**, *131*.

- (103) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.* **2004**, *394*, 334.
- (104) Wheeler, S. E.; Houk, K. N. *J. Chem. Theory Comput.* **2010**, DOI: 10.1021/ct900639.
- (105) Dunlap, B. I. *J. Phys. Chem.* **1986**, *90*, 5524.
- (106) Werpetinski, K. S.; Cook, M. *Phys. Rev. A* **1997**, *52*, R3397.
- (107) Wheeler, S. E.; Houk, K. N. *J. Chem. Theory Comput.* **2009**, *5*, 2301.
- (108) Murray, C. W.; Handy, N. C.; Laming, G. L. *Mol. Phys.* **1993**, *78*, 997–1014.
- (109) Treutler, O. T.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346.
- (110) Wang, X.-G.; Carrington Jr, T. *J. Theor. Comput. Chem.* **2003**, *2*, 599–608.
- (111) Lebedev, V. I. *Zh. Vychisl. Mat. Mat. Fiz.* **1975**, *15*, 48.
- (112) Lebedev, V. I. *Zh. Vychisl. Mat. Mat. Fiz.* **1975**, *16*, 293.
- (113) Kestner, N. R.; Combariza, J. E. In *Reviews in Computational Chemistry*; Boyd, D. B., Lipkowitz, K. B., Eds.; Wiley-VCH: New York, 1999; Vol. 13, pp 99–132.
- (114) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.
- (115) Paizs, B.; Suhai, S. *J. Comput. Chem.* **1998**, *19*, 575–584.

CT100061F

A Comparison of Three Variants of the Generalized Davidson Algorithm for the Partial Diagonalization of Large Non-Hermitian Matrices

Marco Caricato,* Gary W. Trucks, and Michael J. Frisch

Gaussian, Inc., 340 Quinnipiac Street, Bldg. 40,
Wallingford, Connecticut 06492

Received February 24, 2010

Abstract: The solution of the equation of motion coupled cluster singles and doubles problem, that is finding the lowest lying electronic transition energies and properties, is fundamentally a large non-Hermitian matrix diagonalization problem. We implemented and compared three variants of the widely diffuse generalized Davidson algorithm, which iteratively finds the lowest eigenvalues and eigenvectors of such a matrix. Our numerical tests, based on different molecular systems, basis sets, state symmetries, and reference functions, demonstrate that the separate evaluation of the left- and right-hand eigenvectors is the most efficient strategy to solve this problem considering storage, numerical stability, and convergence rate.

1. Introduction

This work originates from our attempts to implement the most efficient algorithm to find the lowest eigenvalues of the approximate Hamiltonian in the equation of motion coupled cluster singles and doubles method (EOM-CCSD).¹ The latter is one of the most accurate and yet affordable methods for the calculation of one-electron transition energies and properties. Its basic equation for the k -th excited state can be written as:

$$(\bar{H}R_k)_c|\Phi_0\rangle = \omega_k R_k|\Phi_0\rangle \quad (1)$$

where $\bar{H} = e^{-T}He^T$ is the similarity transformed Hamiltonian, R_k is an excitation operator toward the k -th state, Φ_0 is the reference function, ω_k is the transition energy, and the notation $(\dots)_c$ indicates that only connected diagrams are considered. In principle, this equation can be solved directly by diagonalization of \bar{H} , whose eigenvalues are the transition energies, and the eigenvectors are R_k , for all the states. In practice, this is not possible because the matrix dimension, roughly o^2v^2 , with o and v being the numbers of occupied and virtual orbitals, is very large. Another complication is that the similarity transformed Hamiltonian is not Hermitian and, hence, has different left and right eigenvectors:

$$\langle\Phi_0|L_k\bar{H} = \omega_k\langle\Phi_0|L_k \quad (2)$$

If the transition energies are of interest, then only eq 1 (or equivalently eq 2) need be solved. However, both left and right eigenvectors are necessary in order to obtain transition properties. For example, the dipole strength between the ground and the k -th excited state is

$$\langle\Phi_0|L_0\mu R_k|\Phi_0\rangle\langle\Phi_0|L_k\mu R_0|\Phi_0\rangle \quad (3)$$

Since the dimension of the \bar{H} matrix prevents a direct diagonalization, the most effective computational approach is a modified version of the Davidson algorithm to treat non-Hermitian matrices.^{2–4} A brief overview of the algorithm is presented in Section 2. However, the basic concept is that the eigenvalues and eigenvectors of interest are obtained through an iterative procedure which avoids the computation, storage, and diagonalization of the complete matrix and stops when certain criteria of convergence are satisfied.

In this work, we compare three variants of the algorithm in ref 4 for the evaluation of the first k eigenvalues and left and right eigenvectors. We implemented these in the Gaussian 09 suite of programs.⁵ In the first two variants, the left and right eigenvectors are converged simultaneously, in the first case expanding them in two sets of biorthonormal trial vectors and in the second expanding them in one set of orthonormal vectors. In the third variant, the right eigen-

* Corresponding author. E-mail: marco@gaussian.com.

vectors are found first and then the left eigenvectors. The results in Section 3 show that the last variant is the most efficient in terms of storage, convergence rate, and numerical stability.

2. Algorithms Description

The algorithms presented in this section can be applied to any non-Hermitian square matrix A of dimension n . Let us start with the algorithm that finds the left and right eigenvectors simultaneously by using two sets of biorthonormal expansion vectors \bar{B} and B , which satisfy:

$$\bar{B}^\dagger B = I \quad (4)$$

where I is the unit matrix. We want to find the lowest k eigenvalues (ω_k) and left (L_k) and right (R_k) eigenvectors of the A matrix, where $k \ll n$. We start with i initial \bar{B} and B vectors (which can be the same) as a guess for the eigenvectors, with $i \geq k$. A is projected onto the subspace of dimension i :

$$A^i = \bar{B}^{i\dagger} A B^i \quad (5)$$

where the superscript i indicates the number of initial vectors. The projected matrix A^i can be diagonalized with standard techniques (its dimension is small), and the eigenvalues and eigenvectors for the subspace are found: ω_k^i , l_k^i , and r_k^i , producing approximate eigenvectors:

$$\begin{aligned} L_k &\simeq L_k^{(i)} = \bar{B}^i l_k^i \\ R_k &\simeq R_k^{(i)} = B^i r_k^i \end{aligned} \quad (6)$$

This approximation can be tested by checking the norm of the residual vectors \bar{W}_k and W_k :

$$\begin{aligned} (A - \omega_k^i) R_k^{(i)} &= W_k \\ (A^\dagger - \omega_k^i) L_k^{(i)} &= \bar{W}_k \end{aligned} \quad (7)$$

Convergence is achieved when the norm of the residuals is below a certain threshold ξ ; otherwise, new vectors are added to the expansion space based on the residuals:

$$\begin{aligned} Q^i &= (\omega_k^i - A_D)^{-1} W_k \\ \bar{Q}^i &= (\omega_k^i - A_D)^{-1} \bar{W}_k \end{aligned} \quad (8)$$

where A_D are the diagonal elements of the A matrix.² This is a good guess for a new set of vectors as long as A is diagonally dominant. This is typically the case for the similarity transformed Hamiltonian.

The vectors \bar{Q}^i and Q^i are then biorthonormalized among each other and with respect to the previous vectors \bar{B}^i and B^i . A set of expansion vectors is thus created, \bar{B}^{2i} and B^{2i} , twice as large as the initial one. With this new set, the matrix A is projected onto a larger subspace:

$$A^{2i} = (\bar{B}^{2i})^\dagger A B^{2i} \quad (9)$$

and the whole process is repeated until convergence. Since the eigenvectors do not converge all at the same rate, a smaller number of new vectors can be created in later iterations as more and more roots converge. More impor-

tantly, this algorithm avoids the explicit calculation and storage of the A matrix, since only products AB and $A^\dagger \bar{B}$ are necessary.⁶

Although this algorithm seems a straightforward extension of the Davidson algorithm, it may encounter numerical instabilities. For instance, complex eigenvalues can be found in intermediate steps, even if the final eigenvalues are real.⁷ A robust way to deal with this issue and to eliminate the intermediate complex eigenvalues in the following cycles is to create twice the number of W and \bar{W} (and thus Q and \bar{Q}) vectors for these eigenvalues, one for the real and one for the imaginary parts, which share the same $R_k^{(i)}$ and $L_k^{(i)}$, eqs 7 and 8. We also found that it is more stable to create two distinguished projected matrices A^i and $A^{i\dagger}$ and to diagonalize them separately in order to evaluate the eigenvectors in the subspace r_k^i and l_k^i (note that the relation $A^i = (A^{i\dagger})^\dagger$ is not exactly satisfied for numerical reasons). This does not add much to the computational time, since the matrix–vector products AB^i and $A^\dagger \bar{B}^i$ are needed anyway for the calculation of the residuals in eq 7, and the final projection has $O(o^2 v^2)$ cost, which is much cheaper than the evaluation of AB^i and $A^\dagger \bar{B}^i$ (which scales as $O(o^2 v^4 + o^3 v^3)$).

Another source of instability arises when approaching convergence. At this point, the right eigenvector for one state may satisfy the convergence criterion, while the corresponding left eigenvector does not or vice versa. One might think that the convergence criteria must be satisfied by the vectors in both spaces before interrupting the creation of new vectors. However, we have found that very small residuals can generate noise that may prevent the convergence of the algorithm. A more robust strategy is to take all the new vectors, Q^i and \bar{Q}^i , and biorthonormalize them with respect to the left space, \bar{B}^i , and right space, B^i , expansion vectors and then among themselves. In this way, the same number of new vectors is created for both spaces, which limits the noise. The same strategy can be used, for example, when a complex eigenvalue is found for one space but not for the other or when the diagonalization of the projected matrices gives eigenvalues that differ more than a certain ratio. Although rare, the latter situation may happen due to the numerical precision of the various operations during the iterative cycles. In the following, we shall refer to this algorithm as “B-Biorth”.

An alternative to the previous algorithm is to orthonormalize the Q^i and \bar{Q}^i with respect to the previous series of B^i and \bar{B}^i and with respect to each other. Since for the first guess $\bar{B}^i = B^i$, the same set of expansion vectors is used for both spaces for all iterations. We shall call this variant “B-Orth”. Note that twice the number of vectors is created for both spaces at each iteration for B-Orth in comparison to B-Biorth. For symmetric matrices, it is known that creating a larger number of expansion vectors than number of target roots helps to increase the convergence rate.⁸ An advantage of this variant compared to the previous algorithm is that it is intrinsically more stable, since the same number of new vectors is created at each iteration for both spaces. However, the possibility of complex eigenvalues at intermediate steps holds, and it is dealt with the same strategy as above.

A third variant, that we shall refer to as “B-1Space” is to evaluate the eigenvectors separately. In this way, only one set of orthonormal B^i vectors is used. This approach can encounter intermediate complex eigenvalues, but does not raise the issue of unbalanced description of the subspaces, as they are spanned separately. Furthermore, the second diagonalization can start with the converged eigenvectors from the first diagonalization, which is usually much better than the initial CIS-based guess. The B-1Space variant does not guarantee that the same eigenvalues are found in the two separate diagonalizations, and a check is necessary after convergence in both spaces. A final biorthonormalization of the converged eigenvectors is performed once the equivalence of both spaces is verified.

For all the algorithms, a maximum subspace dimension can be set according to the details of the calculation and the machine setup. If convergence is not achieved before the subspace limit is reached, then the diagonalization can be restarted by using the last updated eigenvectors as a starting guess. We choose a limit equal to $20 \times n_{\text{states}}$, where n_{states} is the number of states (eigenvalues) to be computed.

If one set of eigenvectors is found, say the right-hand ones and the corresponding eigenvalues, then the left eigenvectors can be also evaluated by solving a linear equation for each root k , eq 2.¹ Standard iterative algorithms can be used to solve these equations.^{1,9} As for EOM-CCSD, eq 2 is very similar to the ground-state Λ vector equation of gradient theory:^{10,11}

$$\langle \Phi_0 | (1 + \Lambda) \bar{H} = 0 \quad (10)$$

in the sense that the same matrix–vector products are involved. Technically, eq 10 is an $Ax = b$ problem with b corresponding to the $\langle ij||ab \rangle$ integrals, whereas eq 2 is an $A'x = 0$ problem; nevertheless, the algorithms employed in the solution of both problems are similar. We shall refer to this alternative as “B-LinSys”. The cost of each step of the iterative diagonalization and the iterative solution of the linear system is comparable. The matrix–vector product $A^\dagger \bar{B}^i$ is the same, and the extra work (evaluation of the residuals in B-1Space) is much cheaper than the building of $A^\dagger \bar{B}^i$. Thus, the difference in efficiency of the two approaches arises from the different rate of convergence of the iterations.

B-Biorth requires the largest amount of storage with two sets of matrices used for the two spaces: \bar{B} and B , $A^\dagger \bar{B}$ and AB , \bar{W} and W , and \bar{Q} and Q . The largest of those are \bar{B} , B , $A^\dagger \bar{B}$, and AB , given that the number of vectors is as large as the maximum dimension of the subspace. \bar{W} , W , \bar{Q} , and Q only require a number of vectors corresponding to the number of target roots. B-Orth is much less demanding, as only one set of B vectors is necessary. B-1Space is the least demanding, as only one set of B , AB , W , and Q matrices are necessary, and the solution of the left-hand problem reuses the same storage. B-LinSys is equivalent to B-1Space. Both storage requirement and computational time can be reduced by exploitation of the equivalence of α and β electrons for closed shell calculations and of Abelian molecular point group symmetry.

The recommended convergence criteria found in the literature is that the norm of the residual vectors must be

below a certain threshold ξ . For EOM-CCSD, Stanton and Bartlett¹ proposed that $\xi = 10^{-5}$ is sufficient to obtain convergence for transition properties. We prefer to use slightly more conservative criteria, thus we check: (i) the norm of the residual vectors; (ii) the change in the eigenvalues ($<\xi \times 10^{-2}$); and (iii) the absolute change in the current eigenvectors ($<\xi$).

3. Results

We report results for four molecular systems: formaldehyde (C_{2v}), ethene (D_{2h}), acetone (C_{2v}), and *trans*-1,3-butadiene (C_{2h}) and six basis sets: 6-31G*, 6-31+G*, 6-31++G**, 6-311++G**, aug-cc-pVDZ, and aug-cc-pVTZ. We computed 3 states for each irrep, thus 12 states for formaldehyde, acetone, and butadiene and 24 for ethene. Restricted closed shell and unrestricted open shell (with a +1 charge) Hartree–Fock (HF) wave functions were considered as reference functions. This range of options allows to test the behavior of the algorithms in a variety of different conditions. The geometries of all the systems were optimized at MP2/6-311+G** level of theory and used for all the excited state calculations.¹²

The results for the closed shell calculations are reported in Table 1. The B-1Space algorithm requires the smallest number of matrix–vector products in all cases, in part because of the use of the converged right eigenvectors as a starting guess for the left eigenvectors.

B-Orth is the least-efficient algorithm, since orthonormalizing the new vectors for both spaces in order to create a single set of vectors seems not to help the convergence. The iterative procedure is also restarted several times, as the expansion subspace is more quickly filled; this further increases the number of cycles necessary to reach convergence and is the reason why in many cases the number of matrix–vector products is more than twice as large as the B-1Space variant. A fairer comparison would be to use a subspace limit twice as large as for the other two algorithms, since twice the number of vectors is added at each cycle, but B-Orth would still be the least efficient choice, and the storage requirement would become even larger than for B-Biorth.

B-Biorth is much closer to B-1Space for the right-hand diagonalization than B-Orth. A slightly larger number of iterations is usually required even in well-behaved cases, since close to convergence some vectors in one space can satisfy the convergence criteria, while the corresponding ones in the other space are slightly off for numerical reasons. Thus, a step where the new vectors of both spaces are orthonormalized to each other is performed in order to add the same number of vectors to both spaces as discussed in Section 2. A larger number of vectors than for B-1Space is, on the other hand, always necessary for the left-hand diagonalization, since the same starting point (CIS eigenvectors) is used for both spaces for B-Biorth. Furthermore, numerical instabilities prevented the convergence for the A_1 irrep of acetone with the 6-31+G* basis set and for the B_u irrep of butadiene with the 6-311++G** basis set. Such instabilities arose close to convergence when numerical noise in the creation of new

Table 1. Number of Matrix–Vector Products for the Ground State (GS) and for Right + Left Eigenvectors with the Three Variants of the Diagonalization Algorithm^a

	GS	B-1Space	B-Orth	B-Biorth	GS	B-1Space	B-Orth	B-Biorth
			6-31 G*				6-31+G*	
formaldehyde (12)	14	159 + 152	323 + 323	214 + 214	15	178 + 157	364 + 364	429 + 429
ethene (24)	12	292 + 252	579 + 579	308 + 308	13	299 + 254	615 + 615	314 + 314
acetone (12)	16	199 + 176	451 + 451	205 + 205	16	208 + 171	467 + 467	nc + 156
butadiene (12)	17	204 + 179	485 + 485	606 + 606	17	201 + 172	449 + 449	203 + 203
			6-31+++G**				6-311+++G**	
formaldehyde (12)	15	174 + 147	357 + 357	177 + 177	15	174 + 152	363 + 363	182 + 182
ethene (24)	13	324 + 267	660 + 660	335 + 335	13	323 + 266	665 + 665	327 + 327
acetone (12)	16	218 + 177	468 + 468	221 + 221	17	220 + 179	492 + 492	220 + 220
butadiene (12)	17	208 + 168	478 + 478	211 + 211	17	206 + 173	491 + 491	149 + nc
			aug-cc-pVDZ				aug-cc-pVTZ	
formaldehyde (12)	15	178 + 159	388 + 388	183 + 183	15	190 + 165	420 + 420	210 + 210
ethene (24)	13	339 + 289	733 + 733	355 + 355	13	347 + 286	734 + 734	348 + 348
acetone (12)	17	227 + 180	502 + 502	228 + 228	17	228 + 188	521 + 521	515 + 515
butadiene (12)	17	215 + 173	523 + 523	225 + 225	17	208 + 165	486 + 486	213 + 213

^a Reference wave function is the restricted HF. Total number of excited states is indicated in parentheses next to the molecule, and nc indicates that the diagonalization is not converged for one of the irreps.

Table 2. Number of Matrix–Vector Products for the Ground State (GS) and for the Right + Left Eigenvectors with the Three Variants of the Diagonalization Algorithm^a

	GS	B-1Space	B-Orth	B-Biorth	GS	B-1Space	B-Orth	B-Biorth
			6-31 G*				6-31+G*	
formaldehyde (12)	17	169 + 154	366 + 366	190 + 190	17	183 + 156	390 + 390	219 + 219
ethene (24)	11	370 + 283	846 + 846	nc+303	12	357 + 299	775 + 775	376 + 376
acetone (12)	22	181 + 152	409 + 409	215 + 215	22	182 + 153	416 + 416	198 + 198
butadiene (12)	17	166 + 141	356 + 356	173 + 173	17	171 + 143	369 + 369	171 + 171
			6-31+++G**				6-311+++G**	
formaldehyde (12)	17	182 + 155	394 + 394	229 + 229	17	180 + 157	395 + 395	202 + 202
ethene (24)	12	367 + 301	776 + 776	374 + 374	12	360 + 298	776 + 776	368 + 368
acetone (12)	22	185 + 153	420 + 420	191 + 191	22	185 + 156	434 + 434	191 + 191
butadiene (12)	17	172 + 143	369 + 369	175 + 175	17	172 + 145	373 + 373	174 + 174
			aug-cc-pVDZ				aug-cc-pVTZ	
formaldehyde (12)	17	182 + 157	393 + 393	194 + 194	19	179 + 159	394 + 394	206 + 206
ethene (24)	12	367 + 299	790 + 790	377 + 377	13	356 + 293	782 + 782	365 + 365
acetone (12)	22	184 + 155	433 + 433	703 + 703	24	183 + 159	438 + 438	nc + 138
butadiene (12)	17	172 + 144	373 + 373	177 + 177	18	175 + 147	374 + 374	176 + 176

^a Reference wave function is the unrestricted HF, and the total charge for each molecule is +1. Total number of excited states is indicated in parentheses next to the molecule, and nc indicates that the diagonalization is not converged for one of the irreps.

vectors unevenly propagated in the two subspaces and the algorithm failed in correcting this effect.

The results for the open shell calculations are reported in Table 2. The same trend as in the closed shell case is observed. We note that the B-Biorth algorithm showed numerical instability for the A_g irrep of ethene with the 6-31G* basis and for the A_2 irrep of acetone with the aug-cc-pVTZ basis (which did not converge). Tables 1 and 2 also report the number of cycles necessary for the convergence of the ground state (GS) CCSD equations. These numbers are close to the average number of cycles per state for the EOM-CCSD equations at least for the right space with the B-1Space algorithm, although the convergence criterion for the ground state equations is one order of magnitude tighter than for the excited state calculation. This is due to the increasing difficulty in converging higher states.

As mentioned in Section 2, when the solution for the two spaces is sought separately, the left eigenvectors can be found by solving a linear system of equations (B-LinSys algorithm) once the eigenvalues for the right space are found with the diagonalization. The number of matrix–vector products for this case are reported in Table 3. A larger number of vectors

Table 3. Number of Matrix–Vector Products for the Left Eigenvectors with the B-LinSys Algorithm for the Closed Shell Case

	6-31 G*	6-31+G*	6-31+++G**	6-311+++G**	aug-cc-pVDZ
formaldehyde	193	215	211	212	223
ethene	342	360	384	379	396
acetone	197	213	232	230	239
butadiene	204	197	209	210	218

than for B-1Space is required. However, the convergence criterion for this algorithm is the root-mean-square (rms) of the norm of the new vectors created in the orthogonalization step ($<\xi \times 10^{-4}$). This is the same as the orthogonalization step for the diagonalization algorithms, but since in B-LinSys no residuals are created at each cycle, there is no other criteria to decide when convergence is reached. In order to test the efficiency of B-LinSys, we reduced the threshold for the formaldehyde case, see Table 4. Although the number of matrix–vector products decreases, the transition properties are not converged for thresholds smaller than $\xi \times 10^{-3}$, with differences of the order of 10^{-4} – 10^{-3} for the oscillator

Table 4. Number of Matrix–Vector Products for the Left Eigenvectors with the B-LinSys Algorithm for the Closed Shell Formaldehyde by Changing the Convergence Threshold ($\xi = 10^{-5}$) in the Orthonormalization of the New Vectors

	6-31 G*	6-31+G*	6-31++G**	6-311++G**	aug-cc-pVDZ
$\xi \times 10^{-4}$	193	215	211	212	223
$\xi \times 10^{-3}$	170	187	187	187	190
$\xi \times 10^{-2}$	146	155	157	154	153
$\xi \times 10^{-1}$	122	123	122	124	125
$\xi \times 10^0$	95	87	87	86	94

strength. With a threshold of $\xi \times 10^{-3}$, B-1Space and B-LinSys are basically equivalent. Therefore, there seems not to be a particular advantage in using the solution of the system of equations over the diagonalization. The latter choice requires the construction of the residuals at each cycle, but this is $O(o^2v^2)$ work for EOM-CCSD and negligible compared to the $A^\dagger \bar{B}^i$ work, which scales as $O(o^2v^4 + o^3v^3)$. For the reasons above and for the practical advantage of using the same code for both spaces, we prefer the B-1Space algorithm to the B-LinSys one.

4. Conclusions

In this paper we present a comparison of three variants of the generalized Davidson algorithm for the iterative diagonalization of large non-Hermitian matrices applied to the EOM-CCSD equations. Two variants seek the right and left eigenvectors simultaneously by using one set of orthonormal trial vectors, B-Orth, or two sets of biorthonormal trial vectors, B-Biorth. A third variant, B-1Space, diagonalizes the matrix from both sides separately and biorthonormalizes the final eigenvectors.

Our numerical tests indicate that the three variants provide the same final results (EOM-CCSD transition energies and properties). The B-1Space option is the most efficient in terms of storage, numerical stability, and convergence rate. The same trend is consistently obtained by varying molecular system, basis set, and reference function (restricted or unrestricted HF). Therefore, the separate left- and right-hand iterative diagonalization is the preferred strategy to find the lowest eigenvalues and eigenvectors of a large non-Hermitian matrix.

Supporting Information Available: Transition energies and oscillator strengths for all the systems are reported in Tables 1–8. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029–7039.
- (2) Davidson, E. R. *J. Comput. Phys.* **1975**, *17*, 87–94.
- (3) Rettrup, S. *J. Comput. Phys.* **1982**, *45*, 100–107.
- (4) Hirao, K.; Nakatsui, H. *J. Comput. Phys.* **1982**, *45*, 246–254.
- (5) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Norm, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, 2009.
- (6) Some elements of the A matrix are never explicitly calculated because the terms in the matrix–vector contractions can be properly organized in order to maintain $O(N^6)$ scaling, where N is the number of basis functions, and storage of at most four indexes quantities.¹
- (7) Here we assume that the target roots of the similarity transformed Hamiltonian are real. Although a generic non-Hermitian matrix may have complex eigenvalues, the above assumption is justified in the context of the EOM-CCSD method because the eigenvalues are excitation energies, which are real quantities. If complex excitation energies are found among the target roots, this is an indication that there is a problem with the description of the wave function (for example the reference function may not be stable), and the rate of convergence of the diagonalization algorithm is, therefore, not relevant.
- (8) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Phys. Chem.* **1998**, *109*, 8218–8224.
- (9) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. In *Numerical Recipes in Fortran*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 1992; pp 22–98.
- (10) Salter, E. A.; Trucks, G. W.; Bartlett, R. J. *J. Chem. Phys.* **1989**, *90*, 1752–1766.
- (11) Gauss, J.; Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1991**, *95*, 2623–2638.
- (12) Caricato, M.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Theory Comput.* **2010**, *6*, 370–383.

CT100111W

JCTC

Journal of Chemical Theory and Computation

Arbitrary-Order Density Functional Response Theory from Automatic Differentiation

Ulf Ekström,^{*,†} Lucas Visscher,[†] Radovan Bast,[‡] Andreas J. Thorvaldsen,[‡] and Kenneth Ruud[‡]

Division of Theoretical Chemistry, Amsterdam Center for Multiscale Modeling, VU University - Faculty of Sciences, De Boelelaan 1083, NL-1081 HV Amsterdam, The Netherlands, and Centre for Theoretical and Computational Chemistry, Department of Chemistry, University of Tromsø, N-9037 Tromsø, Norway

Received March 2, 2010

Abstract: We demonstrate how the functional derivatives appearing in perturbative time-dependent density functional theory can be calculated using automatic differentiation. The approach starts from a computer implementation of the exchange-correlation energy functional, from which arbitrary-order derivatives are generated automatically. Automatic differentiation is shown to provide an accurate, general, and efficient implementation of higher-order exchange-correlation functional derivatives that is easy to maintain. When used in combination with an arbitrary-order response solver, the methodology allows us to generate arbitrary-order response functions from time-dependent density functional theory.

1. Introduction

Time-dependent density functional theory^{1–4} (TDDFT) has become a popular method for the calculation of excited states, due to its low computational cost and, for certain types of excitations, high accuracy. This is despite the fact that present day TDDFT approximations suffer from well-known deficiencies, e.g., in the description of Rydberg excitations, the underestimation of charge-transfer excitations that are associated with significant density transfer, the failure to access explicit many-electron excitations in the adiabatic approximation, and the overestimation of the dynamic polarizability in conducting polymers due to the locality of standard exchange-correlation (XC) kernels. Significant improvements must be made to the approximate energy functionals before TDDFT becomes as reliable for excited states as DFT is for the ground state.⁵ We will here consider response theory, i.e., time-dependent and time-independent perturbation theory based on DFT. We use the term TDDFT to refer to all such approaches, even if the perturbations under consideration may not always be time-dependent. Linear

response TDDFT is the simplest and most commonly used method, but the perturbation theory can be carried out to arbitrary order in the perturbation strengths. An application where the present day functionals may be safely used in conjunction with high order perturbation theory is in the calculation of geometric derivatives. In this case, the perturbation expansion converges (within its radius of convergence) to the same DFT ground state potential energy surface that may be obtained from separate DFT calculations at different molecular geometries.

Several implementations of response theory within TDDFT have been reported,^{6–14} and the calculation of electronic excitation energies and lower-order molecular electromagnetic properties has become routine.^{15–18} In this paper, we discuss the extension of analytic response theory within TDDFT to (in principle) arbitrary order in the Taylor expansion of the energy (or the quasi-energy in the Floquet formalism¹⁹) with respect to field amplitudes (perturbation strengths). This generalization makes it possible to study a wealth of nonlinear optical properties and spectroscopic parameters with (approximate) inclusion of electron correlation effects.

The implementation of TDDFT is technically challenging since, in contrast to the Hartree–Fock exchange matrix within time-dependent Hartree–Fock theory, the expansion

* To whom correspondence should be addressed: E-mail: ulfek@few.vu.nl.

[†] VU University.

[‡] University of Tromsø.

of the XC potential in orders of the density variables does not vanish for second- and higher-order corrections. Repeated application of the chain rule leads therefore to the need for evaluating a large number of derivatives: XC functional derivatives with respect to density variables and derivatives of density variables with respect to the perturbing field strengths. In addition, we may have geometric and magnetic derivatives of basis functions for perturbations which modify the overlap of basis functions such as geometric displacements²⁰ or magnetic perturbations with London atomic orbitals.^{21–23}

The XC energy density derivatives are needed to the same order as the order of the (quasi-) energy derivative. Since most of the modern XC functionals involve rather complicated expressions, higher-order derivatives are often out of reach for manual differentiation. The required derivatives can alternatively be generated using symbolic differentiation techniques (see, e.g., refs 24 and 25), but this approach still requires manual intervention and verification and typically generates rather lengthy code (since every necessary derivative is implemented separately for each functional). In practice, the computer algebra systems do not take issues related to numerical stability into account and may “optimize” statements into numerically unstable forms, something that we will return to in section 4.

Generating functional derivatives via automated symbolic manipulation is therefore not likely to be a practical approach for calculating higher-order XC contributions, although implementations of response functions based on this scheme have been presented in the literature.²⁶

We have for this reason adopted a different approach for calculating higher-order XC energy density derivatives based on automatic differentiation (AD).^{27,28} The basic idea of AD is that every computer program, no matter how complicated, performs a (possibly long) series of simple operations: these are the usual arithmetic operations, together with a small number of intrinsic mathematical functions for exponentials, logarithms, trigonometric functions, etc. AD then makes use of the fact that a computer implementation of an analytical function f contains all information needed for the calculation of derivatives of f , to arbitrary order. This calculation can either be done by taking the source code implementing f as input, and automatically generating the code for the derivative f' , or can be done using operator overloading features of the programming language itself. In both cases, f' is generated from an *existing* implementation of f without manual intervention.

AD has typically been applied to calculate low-order (first and second) derivatives of models with a large number of variables (see for example the applications described in ref 29). In the field of quantum chemistry, we note the application of AD to the calculation of molecular gradients for semiempirical wave functions,³⁰ an application where a large number of low-order derivatives are calculated.

For the present application, the situation is different: we need high-order derivatives of a large number of functions of a few variables. For example, to calculate the cubic response function using a GGA functional, we need to evaluate fourth-order derivatives of the XC energy density

$$\varepsilon_{xc}(\mathbf{r}) = \varepsilon_{xc}(n_{\alpha}(\mathbf{r}), n_{\beta}(\mathbf{r}), |\nabla n_{\alpha}(\mathbf{r})|^2, |\nabla n_{\beta}(\mathbf{r})|^2, \nabla n_{\alpha}(\mathbf{r}) \cdot \nabla n_{\beta}(\mathbf{r})) \quad (1)$$

a function of five local variables that depend on the spin-up (n_{α}) and spin-down (n_{β}) parts of the number density n and their Cartesian gradients, at about a million different grid points for a medium-sized molecule. These derivatives are then multiplied with products of perturbed density variables and integrated to form derivatives of the XC energy with respect to perturbation field strengths, according to the chain rule as described in section 2.

Because of these unusual requirements, we have implemented an AD library based on operator overloading, optimized for high-order derivatives of a small number of variables. The implementation is described in section 3. Numerical stability and performance are tested in sections 4 and 5, respectively, and results of sample applications are presented in section 6. We give some concluding remarks in section 7.

2. Arbitrary-Order Adiabatic Time-Dependent Density Functional Theory

The atomic orbital-based, arbitrary order adiabatic TDDFT formalism employed in this work has recently been presented by Thorvaldsen et al.³¹ The general formalism needed to accommodate response theory in the Kohn–Sham approach was discussed in ref 31, but this paper did not give explicit expressions for the contributions arising from the derivatives of the XC functionals. For future reference, we will present these explicit expressions here for a closed-shell reference state up to the quartic response functions. In the following discussion of the working equations, we will restrict ourselves to the extension from TDHF to TDDFT and, for the sake of clarity, neglect the spin density contributions in this presentation. Our ansatz is thus the XC energy

$$\varepsilon_{xc}(\mathbf{r}) = \varepsilon_{xc}(n(\mathbf{r}), \nabla n(\mathbf{r}) \cdot \nabla n(\mathbf{r})) = \varepsilon_{xc}(n(\mathbf{r}), Z(\mathbf{r})) \quad (2)$$

where we have introduced the square gradient norm, $Z = \nabla n \cdot \nabla n$, of the density. Compared to eq 1, we can work with the total density n instead of the spin-up and spin-down parts and set the spin density $s = n_{\alpha} - n_{\beta}$ to zero. We will, however, point out where and how additional spin density contributions would show up in the working equations. In the case of electric perturbations for a closed-shell reference, as studied in this paper, spin-density contributions are strictly zero for static perturbations but are in general nonzero if the perturbation is frequency dependent. These expressions can be implemented as a straightforward extension of the method described here. Although neglected in this presentation, the reported library for arbitrary-order XC functional derivatives does implement spin-polarized functionals (and derivatives), applicable both to the spin-unrestricted formalism and the spin-restricted formalism with a spin-polarized response.

The additional terms that appear in the working equations when moving from TDHF to Kohn–Sham TDDFT enter in two specific contributions: the first term appears in the XC contribution to the electronic Hessian during the solution of

the set of linear response equations; the second term arises as an additional XC contribution to the perturbed Fock matrix (or matrices) in the contraction of two-electron integrals with the perturbed density matrix expansion. Both contributions require a numerical integration to form Fock-type matrices in the AO basis, $K_{xc;\kappa\lambda}$, with an integrand that can be expressed in the following computationally advantageous form

$$k_{xc;\kappa\lambda}(\mathbf{r}) = u(\mathbf{r}) \Omega_{\kappa\lambda}(\mathbf{r}) + 2\mathbf{v}(\mathbf{r}) \cdot \nabla \Omega_{\kappa\lambda}(\mathbf{r}) \quad (3)$$

containing the AO distribution, $\Omega_{\kappa\lambda} = \phi_{\kappa}^{\dagger} \phi_{\lambda}$, its Cartesian gradient, $\nabla \Omega_{\kappa\lambda}$, and the scalar and vector prefactors u and \mathbf{v} , respectively, which depend on the chosen XC functional and contain functional derivatives of ϵ_{xc} with respect to n and Z , as well as products of functional derivatives and derivatives of (perturbed) density variables. The factor of 2 appearing in eq 3 comes from a product rule differentiation of Z in the AO representation, with respect to the AO density matrix coefficients. In the unperturbed case, the prefactors u and \mathbf{v} are given by

$$u = d_{1,0} \quad (4)$$

$$\mathbf{v} = d_{0,1} \nabla n \quad (5)$$

where $d_{1,0}$ and $d_{0,1}$ represent first-order XC functional derivatives using the short-hand notation:

$$d_{i,j} = \left(\frac{\partial}{\partial n} \right)^i \left(\frac{\partial}{\partial Z} \right)^j \epsilon_{xc} \quad (6)$$

The compact notation of eqs 4 and 5 will be convenient when giving the expressions for the higher-order contributions. The implementation of the XC contribution in the solution algorithm of linear response equations requires the evaluation of derivatives of the prefactors u and \mathbf{v} with respect to a field amplitude b

$$u^b = d_{1,0}^b \quad (7)$$

$$\mathbf{v}^b = d_{0,1}^b \nabla n + d_{0,1} \nabla n^b \quad (8)$$

where we have introduced the notation $d_{i,j}^b = (d/db) d_{i,j}$ etc., for field-perturbed quantities. $d_{1,0}^b$ and $d_{0,1}^b$ are to be expanded using the chain rule:

$$d_{i,j}^b = d_{i+1,j} n^b + d_{i,j+1} Z^{0,b} \quad (9)$$

where $Z^{0,b}$ (0 meaning unperturbed) is short-hand notation for the dot product of two density gradients: $Z^{a,b} = 2\nabla n^a \cdot \nabla n^b$. This means that the first-order field-perturbed prefactors u^b and \mathbf{v}^b contain first- and second-order functional derivatives as well as first-order perturbed density variables, n^b and $Z^{0,b}$. These working equations for the XC contribution to the linear response functions have been given in the literature numerous times with varying notations.^{6–14} Explicit expressions closest to our implementation can be found in ref 32. Note, however, that the expressions in this reference contain additional contributions due to spin magnetization, which we exclude in this presentation. The higher-order XC contributions to the perturbed Fock matrices can be obtained

in a similar manner by straightforward differentiation using the chain rule. The contributions up to the quartic response functions are given in the Appendix.

3. Automatic Differentiation

Our implementation of automatic differentiation is based on replacing all “scalar” floating point operations with operations acting on finite-order Taylor polynomials with floating point coefficients. Each intrinsic mathematical function (exp, log, sin, cos, etc.) of the programming language is first extended to return not only the function value $f(x)$ for a particular argument x , but also the derivatives $f^{(i)}(x)$ for i up to a given order. The derivatives can typically be evaluated using less computational effort than the function value $f(x)$ itself, as is for example the case for the logarithmic function, where the derivative is simply $1/x$. The first derivative of $\sin(x)$ requires the computation of $\cos(x)$, but for the second derivative we obtain again the factor $\sin(x)$, which does not need to be evaluated twice. Similar simplifications can be made for all intrinsic mathematical functions of common programming languages.

Using this code for calculating Taylor expansions of intrinsic mathematical functions f , we are in a position to evaluate expressions $f(g(x))$, where $f(z)$ and $g(x)$ are analytical functions. This is done by first Taylor expanding $g(x)$ up to the desired order as the finite polynomial $P(x)$. This polynomial is then inserted into the Taylor expansion of $f(z)$ around $z = P(0)$. The resulting polynomial is the Taylor polynomial of the composite function. Similar results are obtained for products of functions, i.e., the Taylor expansion of a product of functions can be obtained by first expanding the two functions to a given order and then multiplying their truncated Taylor polynomials. We note that the obtained Taylor coefficients are numerically exact, as they do not arise from any kind of finite difference approximation.

We have implemented efficient arithmetic on multivariate Taylor polynomials in the C++ programming language. Multiplication is performed with a fixed truncation level in every polynomial operation, so that all intermediate values of compound expressions have the same complexity. Using operator overloading techniques, we can convert existing code, working in floating point arithmetic, to a code that works using Taylor polynomials with floating point coefficients. [Operator overloading means that the programmer can give meaning to programming statements such as $z = x*y$, when x , y , and z are of some user defined type. This technique has long been used to implement matrix algebra, and we here use it for Taylor series arithmetic.] As long as the parent code defines an analytical function as a composite expression of intrinsic mathematical functions and arithmetic operations, we can in this way obtain arbitrary-order derivatives of the function. It may seem that this approach is equivalent to rederiving the derivative formula every time the program is run. This is however not the case, since we only compute derivatives at a single expansion point. This is a much simpler task than deriving an analytical expression valid for derivatives at arbitrary points.

In our C++ implementation, we have used the *template* feature of the language to make the number of variables and

polynomial degree compile-time constants. This makes the code very efficient but limits the applicability of the implementation to problems where the number of variables is known at compile time. This is not a limitation for our present application where the number of variables is set by the DFT functional type (two variables for LDA, five for GGA, etc.). The advantage of using templates is that they allow the compiler to produce much more efficient code when the polynomial sizes are known at compile time. This is particularly important for multivariate polynomial multiplication, which has to be formulated recursively. Here, a naive recursive implementation not using templates was found to be 50 times slower than the template-based code.

3.1. Example. In order to illustrate the operating principles of the approach described above, we will give an example of automatic differentiation applied to the simple Slater exchange functional

$$\varepsilon_x(n) = Cn^{4/3} \quad (10)$$

where $C = -0.93$ is a constant. To better illustrate the principles of the approach, we will evaluate the exchange functional as $\varepsilon_x(n) = C(n^4)^{1/3}$.

Suppose we want to Taylor-expand, to third order, $\varepsilon_x(n)$ at $n_0 = 2.0 a_0^{-3}$. We will then deal with third-order Taylor polynomials throughout the calculation. The procedure starts by introducing a small variation, δn , of the density near the expansion point n_0 . The density n can then be written as a polynomial in δn :

$$n = 2.0 + \delta n \quad (11)$$

$$= 2.0 + 1.0\delta n + 0.0\delta n^2 + 0.0\delta n^3 \quad (12)$$

$$\equiv \boxed{2.0 \ 1.0 \ 0.0 \ 0.0} \quad (13)$$

In our implementation, all polynomials have the same order, in this example, order three, so zero coefficients have been explicitly inserted into n for the quadratic and cubic terms. The box notation shows the array of coefficients stored in computer memory and reminds us that the procedure is fully “numerical” in nature—that is, it works with the numerical values of the derivatives and not with their symbolic expressions. Typically the coefficients will be stored as double precision floating point numbers, but in this example, we use two significant digits in the calculation.

Since we choose to evaluate the exchange function as $(n^4)^{1/3}$, the first two steps will be

$$n = \boxed{2.0 \ 1.0 \ 0.0 \ 0.0} \quad (14)$$

$$n^4 = \boxed{16. \ 32. \ 24. \ 8.0}. \quad (15)$$

Here the fourth-order term in n^4 is not needed (because we want only derivatives up to order three), and it is therefore not computed. When the computer now encounters the expression

$$(n^4)^{1/3} = \boxed{\boxed{16. \ 32. \ 24. \ 8.0}}^{1/3} \quad (16)$$

it will generate a third-order Taylor expansion of the cube root function around the constant term of the argument,

which in this case is 16. Introducing a dummy variable z , and using the basic properties of the cube root function, we obtain

$$(16. + z)^{1/3} = 2.5 + 0.052z - 0.0011z^2 + 0.000038z^3 \quad (17)$$

Now, we insert $z = n^4 - 16. = \boxed{0.0 \ 32. \ 24. \ 8.0}$ into this expansion, to obtain

$$\boxed{\boxed{16. \ 32. \ 24. \ 8.0}}^{1/3} = \boxed{2.5 \ 1.7 \ 0.14 \ -0.016} \quad (18)$$

Multiplying, finally, with the constant C , we get

$$C \cdot \boxed{2.5 \ 1.7 \ 0.14 \ -0.016} = \boxed{-2.3 \ -1.6 \ -0.13 \ 0.014}, \quad (19)$$

where the numbers in the last box are now the Taylor coefficients of the Slater exchange functionals at $n = 2.0a_0^{-3}$. This is what we set out to calculate. The result is exact, except for round-off errors, and no finite difference approximation has been used. We note that, except for the Taylor coefficients of the intrinsic mathematical functions such as the cube root used above, we only need to be able to add, subtract, and multiply Taylor polynomials for the scheme to work. For multivariate functions we deal with multivariate Taylor expansions, but the principle remains the same.

The procedure used above may seem like a rather cumbersome way of differentiating eq 10, where we can immediately write down the expression for the derivative to arbitrary order. For more complicated compound expressions, there is however no simpler way of differentiation than to differentiate its parts and combine them using the chain rule. This is exactly what the AD approach does. It is clear that, in some cases, there may be an overhead associated with using AD as described above. In particular, we do not take advantage of sparsity (coefficients that are known *a priori* to be zero) in the derivatives. Since the XC energy and derivatives are evaluated at a large number of grid points, and these evaluations all share the same sparsity pattern, there is an opportunity to optimize the process further. We leave this as a topic for a future study, since the XC energy and derivative evaluation takes only a small amount of time in a typical TDDFT calculation (cf. Figure 2).

4. Numerical Stability

The numerical stability of the scheme described in section 3 depends on the function being differentiated. Loss of precision most often appears when subtracting two almost equal numbers, and if possible such expressions should be reformulated to avoid cancellation error. However, since the AD library provides highly accurate implementations for both the function value and the derivatives of intrinsic mathematical functions, there is less possibility for a loss of precision compared to code generated by a symbolic algebra package. With a symbolic derivative approach, statements are typically reordered, and the final program bears little resemblance to the input provided by the programmer.

We investigate the numerical stability issue by performing the same calculation in double precision (about 16 decimal

Table 1. Relative Accuracy (“Number of Correct Digits”) of AD Density Functional Derivatives, Defined as $\log_{10}(\langle \epsilon_{xc}^{(N)} \rangle / \Delta \epsilon_{xc}^{(N)})$, Where $\langle \epsilon_{xc}^{(N)} \rangle$ Is the Root Mean Square Average of All Partial Derivatives of Order N , and $\Delta \epsilon_{xc}$ Indicates the Difference between the Values Computed in Double Precision and the Highly Accurate Quad-Double Precision Values

order N	LDA ^a	LSDA ^b	LDA ^c	BLYP ^d
0	16.6	16.6	13.1	16.2
1	16.1	16.0	13.2	15.0
2	15.8	16.3	12.4	14.5
3	15.9	15.1	12.2	14.6
4	15.4	14.7	12.0	14.5
5	15.4	14.5	12.0	13.8

^a Evaluated at $n = 1a_0^{-3}$. ^b At $n = 1a_0^{-3}$, $n_\alpha = n_\beta = 0.5a_0^{-3}$. ^c At $n = 10^{-12}a_0^{-3}$. ^d At $n = 2 \times 10^6 a_0^{-3}$, $|\nabla n|^2 = 10^{19} a_0^{-8}$.

digits) and quad-double (64 digits) precision, using the QD library.³³ By taking the quad-double numbers as a reference, we can study the error of the double-precision results. The evaluations are done for densities on the order of unity, as well as very small densities. For BLYP, we use a density that is typical near the nucleus of a heavy atom, where the gradient correction present in BLYP is expected to play a large role. The results are summarized in Table 1 for the LDA (SVWN5)^{34,35} and BLYP^{36–38} functionals. Using double precision arithmetic, we typically obtain 14–15 correct decimal digits, also for the higher-order derivatives. Since the higher-order derivatives are a result of a large number of arithmetic operations, they suffer from some loss of accuracy. In the present examples, we observe a loss of one to two decimal digits in the fifth-order derivatives compared to the accuracy of the XC energy density itself. In some limiting cases, the method suffers from larger errors, as illustrated by the LDA result of Table 1. The derivatives have in this case been evaluated at a very small density, $n = 10^{-12}a_0^{-3}$, and have in this case a relative accuracy of 12–13 digits. A further loss of accuracy is obtained at even smaller densities, but these errors are less important, because low density regions contribute only very little to the total XC energy in a molecule or solid. For all results presented in section 5, we could safely ignore contributions from grid points with an unperturbed density smaller than $10^{-12}a_0^{-3}$. However, for high-order outer valence properties beyond the properties studied in this work, these contributions may still turn out to become significant.

An issue related to the accuracy of the derivatives for small densities is the problem that the derivatives of a function near a singular point may become very large and cause numerical overflow. The standard density functionals are not differentiable at $n = 0$, which may potentially lead to problems. Taking a close look at the fifth-order partial derivatives of the PBE and BLYP functionals (Figure 1), we see that some partial derivatives are indeed very large. The relative errors are however typically below 10^{-10} even at this high order of derivatives. The few derivatives that have larger relative error all have absolute errors smaller than 10^{-12} , which makes these errors negligible in actual calculations. The BLYP functional suffers less from round-off errors, and here the relative errors are smaller than 10^{-12} .

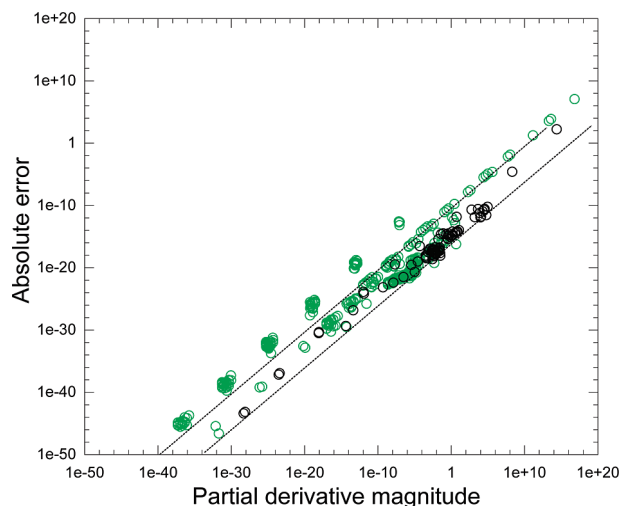


Figure 1. Absolute error as a function of partial derivative magnitude, for the fifth-order derivatives of the BLYP (black) and PBE (green) functionals. Lines are drawn corresponding to relative errors of 10^{-10} (upper line) and 10^{-16} (lower line). The derivatives were evaluated at three different densities: (1.1, 0.9, 17.0, 3.4, 0.1), (1×10^{-6} , 0.9, 1.0, 1×10^3 , 0.1), and (1.0, 0.9, 1×10^6 , 1×10^6 , -1×10^5), using values in atomic units and the variables listed in eq 1.

Locating the exact source of the round-off errors in PBE is left for a future study.

The problem of large derivatives at small densities may be alleviated, if we, instead of expanding $\epsilon_{xc}(n_0 + x)$ in x , expand $\epsilon_{xc}(n_0(1 + x))$, which in effect produces weighted derivatives $n_0^m \epsilon_{xc}^{(m)}(n_0)$. A reciprocal weighting factor n_0^{-m} is introduced in the perturbed density matrices, from which perturbed density variables $n^{b\dots}$ and $Z^{b\dots}$ (eqs 28 and 29 in the Appendix) are calculated, which in both cases prevents numerical over- and underflow.

5. Performance

XC derivatives are rarely a bottleneck in TDDFT calculations, compared to the cost of evaluating the density itself at each gridpoint, but we will nevertheless discuss some performance aspects of our approach and implementation. The computational cost for a given density functional depends on the derivative order N and the number of variables K the functional depends on. For spin-polarized LDA functionals, we have $K = 2$, and for GGA functionals, we have $K = 5$. There are a total of $M_N^K = \binom{K+N}{N} = \mathcal{O}(N^K)$ partial derivatives up to order N . Product expressions, $f(n)g(n)$, are evaluated using “naive” polynomial multiplication, requiring $\mathcal{O}(N^{2K})$ operations. For the evaluation of intrinsic mathematical functions such as $\exp(f(n))$, a total of $\mathcal{O}(N^{2K+1})$ floating point operations are needed for the evaluation of all partial derivatives (although we have in many parts of the implementation used the “fast” algorithms that exist for the manipulation of Taylor series³⁹). We can therefore expect that, for a given XC functional, the asymptotic cost for calculating derivatives up to order N with respect to K variables is $\mathcal{O}(N^{2K+1})$. However, we are rarely interested in the asymptotic behavior since the derivative order N is in practice limited to rather small values. The

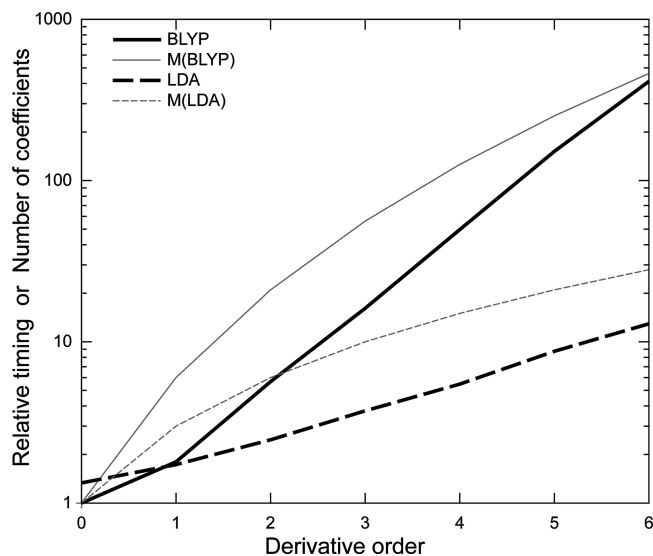


Figure 2. Timings for the evaluation of all partial derivatives of LDA and BLYP functionals up to a given order. Also shown is the number of such partial derivative coefficients, $M(\text{LDA})$ and $M(\text{BLYP})$. Timings are relative to the time used in evaluating the BLYP XC energy. On a rather modest CPU (1.7 GHz Intel Pentium M), this operation takes 3.0 s for 10^6 grid points.

performance for small N values is strongly affected by the particular implementation, as well as compiler optimizations and CPU architecture. We therefore show in Figure 2 real timings for LDA and BLYP derivatives up to order $N = 6$. These timings are plotted together with the M_N^K factor, showing that up to order six the number of partial derivatives grows faster than the time taken to compute them for the BLYP ($K = 5$) and LDA ($K = 2$) functionals. The reason for this is that, for the calculation of derivatives, almost no additional evaluations of intrinsic mathematical functions are needed. From Figure 2, we conclude that, even for a very high order of the functional derivatives, the XC contribution is unlikely to be a computational bottleneck in TDDFT calculations since the number of evaluations grows linearly with the number of grid points which in turn typically grow linearly with the system size.

Table 2. Components of the Static and Frequency-Dependent Polarizability, and the First, Second, and Third Hyperpolarizabilities of FH Calculated Using the q-aug-cc-pVTZ Basis Set^a

	$\omega = 0$		$\omega = 0.06562 \text{ au}$		$\omega = 0.072 \text{ au}$	
	HF	LDA	HF	LDA	HF	LDA
$\alpha(x, x)$	4.495	5.930	4.529	6.013	4.537	6.030
$\alpha(z, z)$	5.759	6.854	5.802	6.924	5.811	6.939
$\beta(x, z, x)$	-0.5087	-2.329	-0.6237	-3.074	-0.6519	-3.274
$\beta(z, x, x)$	-0.5087	-2.329	-0.5106	-2.632	-0.5101	-2.701
$\beta(z, z, z)$	-8.397	-10.52	-9.056	-11.72	-9.200	-11.99
$\gamma(x, x, x, x)$	335.9	1148	429.6	1887	453.9	2140
$\gamma(x, z, z, x)$	96.87	309.6	126.3	549.7	134.3	639.6
$\gamma(z, z, x, x)$	96.87	309.6	118.4	446.2	123.4	484.9
$\gamma(z, z, z, z)$	279.6	636.1	342.3	876.7	357.5	942.7
$\delta(x, z, x, x, x)$	111.3	592.6	-199.5	-11218	-393.6	-31464
$\delta(z, x, x, x, x)$	111.3	592.6	257.6	65.90	328.3	-250.5
$\delta(x, z, z, z, x)$	75.95	1618	-302.2	-7668	-554.5	-30618
$\delta(z, z, z, x, x)$	75.95	1619	-39.63	2465	-79.56	2951
$\delta(z, z, z, z, z)$	-1484.2	-2079	-3062	-8340	-3574	-11374

^a All numbers in atomic units; all perturbing dipole operators carry the same frequency, 0, 0.06562 au, or 0.072 au; $R_e = 1.7328$ bohr; the direction of the positive z axis is from F to H.

6. Results and Discussion

As a demonstration of our approach, we have calculated static and frequency-dependent first, second, and third hyperpolarizabilities, $\beta(-2\omega; \omega, \omega)$, $\gamma(-3\omega; \omega, \omega, \omega)$, and $\delta(-4\omega; \omega, \omega, \omega)$, of the FH molecule. All calculations were performed with a locally modified version of the DIRAC quantum chemistry package,⁴⁰ using the radial quadrature proposed by Lindh et al.,⁴¹ Lebedev grids⁴² for integration on spheres, and (quadruple augmented) cc-pV{D,T}Z basis sets of Dunning and co-workers.⁴³

Selected nonzero components of the calculated (hyper)polarizability tensors obtained using the HF and LDA methods, respectively, are reported in Table 2. Both the static and the frequency-dependent LDA (hyper)polarizabilities are all consistently larger in magnitude than the HF results. This is in agreement with the smaller HOMO–LUMO energy gap of LDA ($0.33 E_h$) compared to HF ($0.65 E_h$). Note that the largest second hyperpolarizability tensor elements (both static and frequency dependent) are the components perpendicular to the molecular axis, $\gamma(x; x, x, x)$ in Table 2, and not the parallel tensor element, $\gamma(z; z, z, z)$. We can also note that, while the static HF and LDA (hyper)polarizabilities have consistent signs, this is not the case for frequency-dependent $\delta(z; x, x, x, x)$ and $\delta(z; z, z, x, x)$ elements.

We would like to mention that no GGA results are reported in Table 2, although we have access to numerically stable analytic arbitrary-order GGA functional derivatives and analytic derivatives of perturbed Kohn–Sham matrix elements. During extensive calibration studies, we have observed that the numerical integration of the GGA γ and δ elements of FH is difficult. Using presently available XC numerical integration grids, we are not able to obtain GGA results of FH that are stable with respect to changes in the grid, and we have therefore omitted these results from the discussion. We emphasize, however, that, for a fixed grid, our results for the GGAs correspond in the static cases to the results obtained from a finite difference of lower-order (hyper)polarizabilities.

In order to illustrate the problem in the integration of higher-order GGA contributions we have plotted the distri-

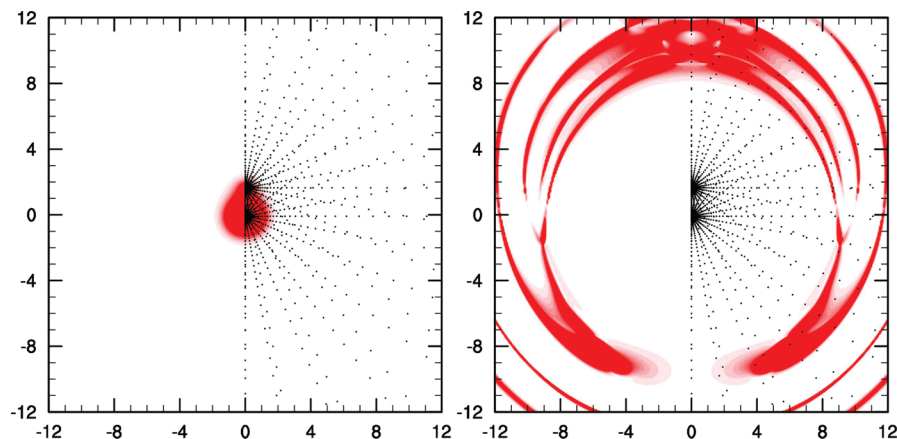


Figure 3. BLYP XC energy density of the FH molecule (left panel) and $|v^{bcd}|$ of eq 26 (right panel; all perturbations are static and parallel to the molecular axis). The color intensity is proportional to the respective absolute value. The numerical integration grid points are represented as dots which “radiate” from the atom centers at (0, 0, 0) bohr and (0, 0, 1.7328) bohr (dimensions: 24×24 bohr).

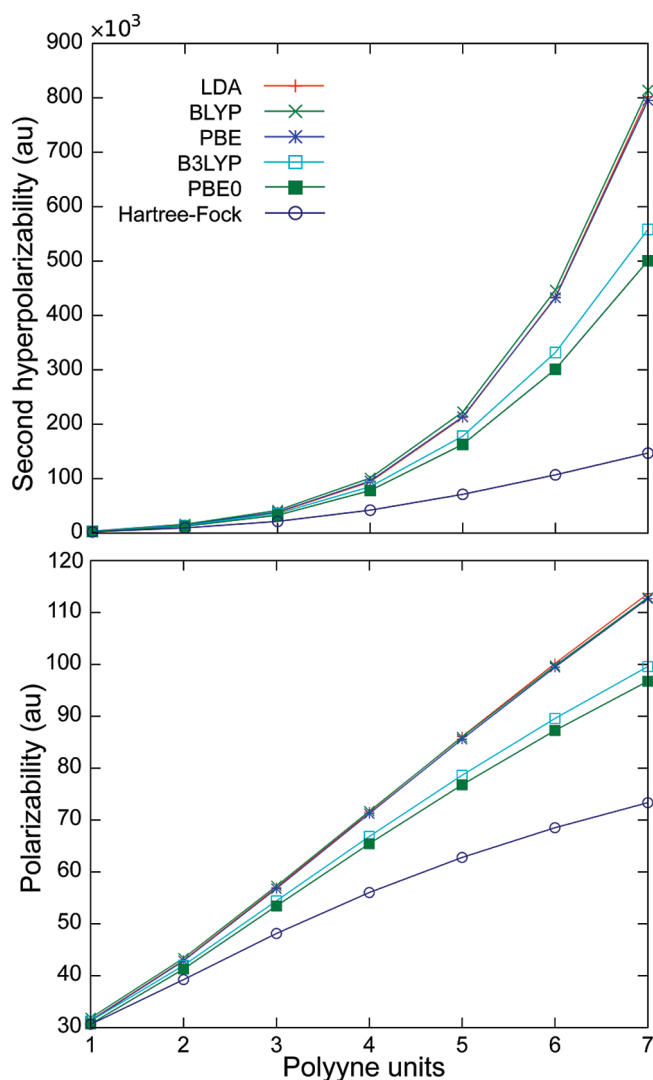


Figure 4. Static polarizability $\alpha(z, z)$ and second hyperpolarizability $\gamma(z, z, z, z)$ per polyene unit for a linear polyene ($C_{2N}H_2$) chain (aug-cc-pVDZ; same geometries as in ref 46). LDA, BLYP, and PBE curves are nearly identical at this scale.

bution of grid points and representative functions in Figure 3. In the left panel of Figure 3, we plot the BLYP XC energy density and observe that it is well represented by a grid which

is dense in the regions where the XC energy density is significant. The situation is very different in the integration of higher-order GGA terms, where the dominating function values—visible as a series of rings in the right panel in Figure 3—are only poorly sampled by the integration grid. We have found that already β calculations may require a new integration strategy in order to obtain fully converged results.

To remedy the illustrated deficiency of the presently available numerical integration grid, we are currently investigating alternative integration techniques. We focus in particular on adaptive numerical integration schemes which would allow for a more flexible and balanced representation of higher-order valence properties as well as mixed geometric-electromagnetic properties such as Raman optical activity.⁴⁴

We emphasize that these problems are not unique to our AD implementation but are inherent to *all* perturbative TDDFT calculations based on eq 3 and probably strongly dependent on the molecule studied. To give an example where the GGA numerical integration was unproblematic, we plot the parallel static polarizability and second hyperpolarizability for linear polyene ($C_{2N}H_2$) chains in Figure 4. These quantities are severely overestimated by nonhybrid XC functionals.⁴⁵ The numerical integration of the parallel component of α and γ was found to be stable to variations in grid parameters with LDA, BLYP, and PBE curves being nearly identical at the scale of Figure 4.

Finally, we would like to mention that the calculation of higher-order valence properties requires not only well calibrated numerical grids but also increasingly diffuse basis sets. Very diffuse basis sets with an even higher augmentation level than the basis sets employed in this work may cause numerical problems due to linear dependencies already in the ground state calculations in addition to a challenging numerical integration of the higher-order XC contributions.

7. Conclusions

We have shown how high-order, time-dependent density functional theory methods can be reliably and efficiently

implemented, using automatic differentiation when evaluating the XC energy derivatives. Numerical roundoff errors are negligible even at very high (fifth) order derivatives.

We have presented HF and LDA static and frequency-dependent first, second, and third hyperpolarizabilities of the FH molecule and discussed the presently challenging numerical integration to obtain the corresponding GGA results. Applying the arbitrary-order response methods to the second hyperpolarizability of polyene chains, we show that this quantity is severely overestimated by the LDA XC functional, and that using GGA functionals beyond the adiabatic LDA approximation does not improve the results. We expect the results to improve using time-dependent current-density-functional theory^{45,47} or with an exact-exchange DFT approach.^{48–50}

To facilitate a more widespread use of the AD method in the DFT community, we have developed a generic software library, XCFun,⁵¹ for calculating arbitrary-order XC derivatives, using the approaches described above. It is similar in scope to the Libxc library⁵² but provides derivatives to arbitrary order and works with any set of density variables (for example, n_α and n_β or $n = n_\alpha + n_\beta$ and $s = n_\alpha - n_\beta$). XCFun is therefore suitable both for development of new XC functionals and for calculations of DFT response properties to arbitrary order.

Acknowledgment. This work has received support from the Norwegian Research Council through a Centre of Excellence Grant (Grant No. 179568/V30), a YFF grant to K.R. (Grant No. 162746/V00), and a VIBRON Grant (Grant No. 177558/V30), as well as through a grant of computer time from the Norwegian Supercomputing Program. L.V. thanks The Netherlands Organisation for Scientific Research (NWO) for support through the VICI programme. U.E. acknowledges support from the Wenner-Gren foundations.

Appendix

A. Spin Density Contribution. If the spin-density contributions also were to be included, eq 3 would contain additional terms, one term, $k_{xc;\kappa\lambda}^z$, in the collinear spin density approximation or three terms, $k_{xc;\kappa\lambda}^x$, $k_{xc;\kappa\lambda}^y$, and $k_{xc;\kappa\lambda}^z$, in the noncollinear spin density approach, where

$$k_{xc;\kappa\lambda}^\mu = u_\mu \Omega_{\kappa\lambda}^\mu + 2\mathbf{v}_\mu \cdot \nabla \Omega_{\kappa\lambda}^\mu \quad (20)$$

and $\Omega_{\kappa\lambda}^\mu = \phi_\kappa^\dagger \sigma_\mu \phi_\lambda$ with σ_μ being one of the Pauli spin matrices ($\mu = x, y, z$). In addition, eq 9 and all higher-order contributions given below would include additional perturbed density variables. Within linear response, corresponding noncollinear spin density contributions can be found for instance in ref 32.

B. Second-Order XC Contribution. The second-order XC contribution requires second-order field-perturbed prefactors u^{bc} and \mathbf{v}^{bc}

$$u^{bc} = d_{1,0}^{bc} \quad (21)$$

$$\mathbf{v}^{bc} = d_{0,1}^{bc} \nabla n + d_{0,1}^b \nabla n^c + d_{0,1}^c \nabla n^b \quad (22)$$

where $d_{1,0}^{bc}$ and $d_{0,1}^{bc}$ are to be expanded using

$$d_{i,j}^{bc} = d_{i+1,j}^b n^c + d_{i,j+1}^b Z^{0,c} + d_{i,j+1}^c Z^{b,c} \quad (23)$$

The lower-order terms $d_{i+1,j}^b$ and $d_{i,j+1}^c$ are to be expanded according to eq 9. This means that the second-order field-perturbed prefactors u^{bc} and \mathbf{v}^{bc} contain first-, second-, and third-order functional derivatives as well as first-order derivatives of density variables, and the second-order term $Z^{b,c} = 2\nabla n^b \cdot \nabla n^c$. Observe that terms containing the highest-order density matrix (here, terms containing n^{bc}) are not present due to the $2n + 1$ rule.

C. Third- and Fourth-Order XC Contributions. Third- and fourth-order prefactors for cubic and quartic response functions, respectively, can be obtained accordingly:

$$u^{bcd} = d_{1,0}^{bcd} \quad (24)$$

$$u^{bcde} = d_{1,0}^{bcde} \quad (25)$$

$$\mathbf{v}^{bcd} = d_{0,1}^{bcd} \nabla n + d_{0,1}^{bc} \nabla n^d + d_{0,1}^{bd} \nabla n^c + d_{0,1}^{cd} \nabla n^b + d_{0,1}^b \nabla n^{cd} + d_{0,1}^c \nabla n^{bd} + d_{0,1}^d \nabla n^{bc} \quad (26)$$

$$\mathbf{v}^{bcde} = d_{0,1}^{bcde} \nabla n + d_{0,1}^{bcd} \nabla n^e + d_{0,1}^{bce} \nabla n^d + d_{0,1}^{bde} \nabla n^c + d_{0,1}^{cde} \nabla n^b + d_{0,1}^{bc} \nabla n^{de} + d_{0,1}^{bd} \nabla n^{ce} + d_{0,1}^{be} \nabla n^{cd} + d_{0,1}^{cd} \nabla n^{be} + d_{0,1}^{ce} \nabla n^{bd} + d_{0,1}^{de} \nabla n^{bc} + d_{0,1}^b \nabla n^{cde} + d_{0,1}^c \nabla n^{bde} + d_{0,1}^d \nabla n^{bce} + d_{0,1}^e \nabla n^{bcd} \quad (27)$$

These terms require the evaluation of a growing number of recursive terms and perturbed density variables. The third-order term reads as

$$d_{i,j}^{bcd} = d_{i+1,j}^{bc} n^d + d_{i+1,j}^b n^{cd} + d_{i+1,j}^c n^{bd} + d_{i+1,j}^d n^{bc} + d_{i,j+1}^{bc} Z^{0,d} + d_{i,j+1}^b (Z^{0,cd} + Z^{c,d}) + d_{i,j+1}^c (Z^{0,bd} + Z^{b,d}) + d_{i,j+1}^d (Z^{0,bc} + Z^{b,c}) + d_{i,j+1}^e (Z^{b,cd} + Z^{c,bd} + Z^{d,bc}) \quad (28)$$

and the fourth-order term can be written as

$$\begin{aligned}
d_{ij}^{bcde} = & d_{i+1,j}^{bcd} n^e \\
& + d_{i+1,j}^{bc} n^{de} + d_{i+1,j}^{bd} n^{ce} + d_{i+1,j}^{be} n^{cd} + d_{i+1,j}^{cd} n^{be} \\
& + d_{i+1,j}^{ce} n^{bd} + d_{i+1,j}^{de} n^{bc} \\
& + d_{i+1,j}^b n^{cde} + d_{i+1,j}^c n^{bde} + d_{i+1,j}^d n^{bce} + d_{i+1,j}^e n^{bcd} \\
& + d_{i,j+1}^{bcd} Z^{0,e} \\
& + d_{i,j+1}^{bc} (Z^{0,de} + Z^{d,e}) \\
& + d_{i,j+1}^{bd} (Z^{0,ce} + Z^{c,e}) \\
& + d_{i,j+1}^{be} (Z^{0,cd} + Z^{c,d}) \\
& + d_{i,j+1}^{cd} (Z^{0,be} + Z^{b,e}) \\
& + d_{i,j+1}^{ce} (Z^{0,bd} + Z^{b,d}) \\
& + d_{i,j+1}^{de} (Z^{0,bc} + Z^{b,c}) \\
& + d_{i,j+1}^b (Z^{0,cde} + Z^{c,de} + Z^{d,ce} + Z^{e,cd}) \\
& + d_{i,j+1}^c (Z^{0,bde} + Z^{b,de} + Z^{d,be} + Z^{e,bd}) \\
& + d_{i,j+1}^d (Z^{0,bce} + Z^{b,ce} + Z^{c,be} + Z^{e,bc}) \\
& + d_{i,j+1}^e (Z^{0,bcd} + Z^{b,cd} + Z^{c,bd} + Z^{d,bc}) \\
& + d_{i,j+1}^b (Z^{b,cde} + Z^{c,bde} + Z^{d,bce} \\
& + Z^{e,bcd} + Z^{b,de} + Z^{bd,ce} + Z^{be,cd})
\end{aligned} \tag{29}$$

Expressions of even higher-order terms as well as the inclusion of spin density variables can be achieved rather straightforwardly using automatic code generation techniques. Note that in this work and in the above discussion we also omit contributions to accommodate perturbations which modify the overlap of basis functions, such as geometric displacements or magnetic perturbations with London atomic orbitals.

References

- Gross, E. K. U.; Kohn, W. *Adv. Quantum Chem.* **1990**, *21*, 255.
- Casida, M. Time-dependent density-functional response theory for molecules. In *Recent Advances in Density Functional methods, Part I*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; p 155.
- van Leeuwen, R. *Int. J. Mod. Phys. B* **2001**, *50*, 1969.
- Marques, M. A. L.; Gross, E. K. U. *Annu. Rev. Phys. Chem.* **2004**, *55*, 427.
- Casida, M. E. *J. Mol. Struct. (Theochem)* **2009**, *914*, 3.
- Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454.
- Jamorski, C.; Casida, M. E.; Salahub, D. R. *J. Chem. Phys.* **1996**, *104*, 5134.
- Petersilka, M.; Gossmann, U. J.; Gross, E. K. U. *Phys. Rev. Lett.* **1996**, *76*, 1212.
- Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218.
- Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180.
- Hirata, S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *302*, 375.
- Görling, A.; Heinze, H. H.; Ruzankin, S. P.; Staufer, M.; Rösch, N. *J. Chem. Phys.* **1999**, *110*, 2785.
- van Gisbergen, S. J. A.; Snijders, J. G.; Baerends, E. J. *Comput. Phys. Commun.* **1999**, *118*, 119.
- Rinkevicius, Z.; Tunell, I.; Sałek, P.; Vahtras, O.; Ågren, H. *J. Chem. Phys.* **2003**, *119*, 34.
- Burke, K.; Werschnik, J.; Gross, E. K. U. *J. Chem. Phys.* **2005**, *123*, 062206.
- Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009.
- Marques, M.; Ullrich, C. A.; Noguiera, F.; Rubio, A.; Burke, K.; Gross, E. K. U. *Time-dependent density functional theory*; Springer: Heidelberg, Germany, 2006.
- Elliott, P.; Furche, F.; Burke, K. Excited states from time-dependent density functional theory. In *Rev. Comput. Chem.*; Lipkowitz, K. B., Cundari, T. R., Eds.; Wiley: Hoboken, NJ, 2009; p 91.
- Christiansen, O.; Jørgensen, P.; Hättig, C. *Int. J. Quantum Chem.* **1997**, *68*, 1.
- Pulay, P. *Mol. Phys.* **1969**, *17*, 197.
- London, F. *J. Phys. Radium* **1937**, *8*, 397.
- Ruud, K.; Helgaker, T.; Bak, K. L.; Jørgensen, P.; Jensen, H. J. Aa. *J. Chem. Phys.* **1993**, *99*, 3847.
- Helgaker, T.; Wilson, P. J.; Amos, R. D.; Handy, N. C. *J. Chem. Phys.* **2000**, *113*, 2983.
- Strange, R.; Manby, F. R.; Knowles, P. J. *Comput. Phys. Commun.* **2001**, *136*, 310.
- Sašek, P.; Hesselmann, A. *J. Comput. Chem.* **2007**, *28*, 2569.
- Jansík, B.; Sašek, P.; Jonsson, D.; Vahtras, O.; Ågren, H. *J. Chem. Phys.* **2005**, *122*, 054107.
- Rall, L. B. *Automatic Differentiation: Techniques and Applications*; Springer: Berlin, 1981; Vol. 120.
- Griewank, A.; Walther, A. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, 2nd ed.; SIAM: Philadelphia, PA, 2008; Other Titles in Applied Mathematics 105.
- Advances in Automatic Differentiation*; Bischof, C. H., Bücker, H. M., Hovland, P. D., Naumann, U., Utke, J., Eds.; Springer: Berlin, 2008; Vol. 64.
- Steiger, R.; Bischof, C.; Lang, B.; Thiel, W. *Future Gen. Comput. Syst.* **2005**, *21*, 1324.
- Thorvaldsen, A. J.; Ruud, K.; Kristensen, K.; Jørgensen, P.; Coriani, S. *J. Chem. Phys.* **2008**, *129*, 214108.
- Bast, R.; Jensen, H. J. Aa.; Saue, T. *Int. J. Quantum Chem.* **2009**, *109*, 2091.
- Hida, Y.; Li, X.; Bailey, D. Quad-Double/Double-Double Computation Package. <http://crd.lbl.gov/dhbailey/mpdist/> (accessed May 2010).
- Slater, J. C. *Phys. Rev.* **1951**, *81*, 385.
- Vosko, S. J.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200.
- Brent, R. P.; Kung, H. T. *J. Assoc. Comput. Machinery* **1978**, *25*, 581.
- Development version of DIRAC, a relativistic ab initio electronic structure program, release DIRAC08 (2008), written by: Visscher, L., Jensen, H. J. Aa., Saue, T., with new contributions from Bast, R., Dubillard, S., Dyall, K. G., Ekström, U., Eliav, E., Fleig, T., Gomes, A. S. P., Helgaker, T. U., Henriksson, J., Iliaš, M., Jacob, Ch. R., Knecht, S.,

- Norman, P., Olsen, J., Pernpointner, M., Ruud, K., Salek, P., Sikkema J. (see <http://dirac.chem.sdu.dk>, accessed May 2010).
- (41) Lindh, R.; Malmqvist, P.-A.; Gagliardi, L. *Theor. Chem. Acc.* **2001**, *106*, 178.
- (42) Lebedev, V. I.; Laikov, D. N. *Doklady Mathematics* **1999**, *59*, 477.
- (43) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Phys. Chem.* **1992**, *96*, 6769.
- (44) Ruud, K.; Thorvaldsen, A. J. *Chirality* **2009**, *21*, S54.
- (45) van Faassen, M.; de Boeij, P. L.; van Leeuwen, R.; Berger, J. A.; Snijders, J. G. *Phys. Rev. Lett.* **2002**, *88*, 186401.
- (46) Dalskov, E. K.; Oddershede, J.; Bishop, D. M. *J. Chem. Phys.* **1999**, *108*, 2152.
- (47) Vignale, G.; Kohn, W. *Phys. Rev. Lett.* **1996**, *77*, 2037.
- (48) Bulat, F. A.; Toro-Labbé, A.; Champagne, B.; Kirtman, B.; Yang, W. *J. Chem. Phys.* **2005**, *123*, 014319.
- (49) Grüning, M.; Gritsenko, O. V.; Baerends, E. J. *J. Chem. Phys.* **2002**, *116*, 6435.
- (50) Karolewski, A.; Armiento, R.; Kümmel, S. *Chem. Phys.* **2009**, *5*, 712.
- (51) Ekström, U. XCFun library. <http://www.admol.org/xcfun> (accessed May 2010).
- (52) Libxc library. <http://www.tddft.org/programs/octopus/wiki/index.php/Libxc> (accessed May 2010).

CT100117S

Approximate Multiconfigurational Treatment of Spin-Coupled Metal Complexes

Guilherme Menegon Arantes*[†] and Peter R. Taylor[‡]

Instituto de Química, Universidade de São Paulo, Av. Lineu Prestes 748, 05508-900, São Paulo, SP, Brazil and Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom

Received March 8, 2010

Abstract: The weak interaction between unpaired electrons in polynuclear transition-metal complexes is often described by exchange and spin polarization mechanisms. The resulting intrinsic multiconfigurational electronic structure for such complexes may be calculated with wave function-based methods (e.g., complete active space configuration interaction and complete active space self-consistent field), but computations become extremely demanding and even unfeasible for polynuclear complexes with a large number of open-shells. Here, several levels of selection of configurations and symmetry considerations that still capture the essential physics of exchange and spin polarization mechanisms are presented. The proposed approximations result in significantly smaller configuration interaction expansions and are equally valid for ab initio and semiempirical methods. Tests are performed in simple molecular systems and in small transition-metal complexes that cover a range of valence and charge states. In particular, superexchange contributions can be calculated to good accuracy using only single ionic excitations. Further reduction in the size of the configuration expansions is possible but restricts the description to low-lying spin ladders. The proposed configuration interaction schemes may be used to resolve space and spin symmetries in the calculation of electronic structures, exchange coupling constants, and other properties pertinent to polynuclear transition-metal complexes.

1. Introduction

Polynuclear transition-metal (TM) compounds with weakly coupled open-shell electrons have interesting magnetic properties as a consequence of the population at thermal energies of low-lying excited states with different total spins. The underlying interactions are traditionally mapped to a spin–spin coupling between momenta \mathbf{S} localized in neighboring magnetic sites and are often described by the Heisenberg–Dirac–van Vleck spin Hamiltonian:¹

$$\hat{H}_{\text{HDvV}} = - \sum_{A < B} J_{AB} \mathbf{S}_A \cdot \mathbf{S}_B \quad (1)$$

where J_{AB} is the isotropic Heisenberg coupling constant between spins on sites A and B. Since $[\hat{H}_{\text{HDvV}}, \hat{S}^2] = 0$, the two operators share a common set of eigenstates. The eigenvalues correspond to a spin ladder, and the energy gaps between low-lying spin states depend linearly on the J coupling constant. For the simplest case of a pair of magnetic sites with spins \mathbf{S}_A and \mathbf{S}_B , the coupling is ferromagnetic, and $J > 0$ in the sign convention assumed on eq 1, if the ground state is high-spin $S = S_A + S_B$. The coupling is antiferromagnetic, and $J < 0$, if the ground state is low-spin $S = |S_A - S_B|$. The spin–spin interaction modeled by eq 1 is in fact an effective one. As proposed by Heisenberg² and Dirac,³ the interactions arise due to spin-independent Coulomb electron–electron repulsion and exchange symmetry.

First-principles calculations with the spin-free electronic Hamiltonian should then be able to predict spin eigenstates and J constants for TM compounds. Anderson^{4,5} was seminal

* Corresponding author. Email: garantes@iq.usp.br.

[†] Universidade de São Paulo.

[‡] University of Warwick.

in realizing how to extract the main contributions to the effective spin coupling from the electronic structure. His model can be understood by considering the following simple valence-bond (VB) scheme: A pair of magnetic sites A and B contain two weakly interacting electrons occupying two orthogonal orbitals (constructed by a suitable rotation of the molecular orbitals) labeled a and b localized on centers A and B, respectively. By weakly interacting it should mean that the two electrons do not form a covalent bond. This situation corresponds to a dihydrogen molecule at stretched bond distance or a spin-coupled Cu(II, d^9) dimer. Four Slater determinants with $M_S = 0$ can be constructed: $|a\bar{b}|$, $|\bar{a}b|$, $|a\bar{a}|$, and $|\bar{b}\bar{b}|$. The first two are “neutral” configurations, and the last two are charge-transfer “ionic” configurations. Their combination results in the following configuration state functions:

$$\begin{aligned} |^1\Psi_{\text{neu}}\rangle &= 2^{-1/2}[|a\bar{b}| - |\bar{a}b|] \\ |^1\Psi_{\text{ion}}^A\rangle &= |a\bar{a}| \\ |^1\Psi_{\text{ion}}^B\rangle &= |b\bar{b}| \\ |^3\Psi_{\text{neu}}\rangle &= 2^{-1/2}[|a\bar{b}| + |\bar{a}b|] \end{aligned} \quad (2)$$

The energy difference between the triplet $|^3\Psi_{\text{neu}}\rangle$ and the singlet $|^1\Psi_{\text{neu}}\rangle$ will be proportional to K_{ab} , the exchange integral between orbitals a and b .⁶ This direct exchange interaction is ferromagnetic because the high-spin state (triplet) is favored. Configuration mixing between neutral and ionic states will lower the singlet energy and lead to the ground state:⁶

$$|^1\Psi_{\text{CI}}\rangle = (1 - \alpha)^{1/2}|^1\Psi_{\text{neu}}\rangle + \alpha^{1/2}|^1\Psi_{\text{ion}}\rangle \quad (3)$$

where $|^1\Psi_{\text{ion}}\rangle$ is a superposition of the two ionic configuration state functions shown above and α gives the degree of mixing between the neutral and ionic states. This mixing is equivalent to a virtual hopping of the electron from one magnetic site to the other (the charge-transfer ionic configurations), and it gives an antiferromagnetic contribution to spin-coupling because the low spin (singlet) is favored. In general, for weakly coupled open-shell compounds with several unpaired electrons, neutral configurations will appear in the wave function expansion for all spin states. Their contribution to spin coupling is ferromagnetic, i.e., stabilize the high-spin state, and is known as the *direct exchange* effect or mechanism. Ionic configurations will appear in expansions of all but the highest spin state and give antiferromagnetic contributions known as the through-space *superexchange* mechanism.

This simple VB model can be expanded to explicitly include an occupied valence closed shell of diamagnetic ligand bridges that coordinate metal ions in TM complexes. Ligand-to-metal charge-transfer (LMCT) excitations built out of a set of neutral and ionic configurations, equivalent to those on eq 3, will have either anti- or ferromagnetic contributions to spin coupling. This issue is discussed in more detail below. To make a connection with the jargon of previous perturbative treatments,^{7–9} it should be noted that single LMCT excitations out of neutral configurations are usually called ligand spin polarization (LSP) because an

effective spin density appears on the bridge.⁸ Double LMCT excitations are termed dynamic or double spin polarization (DSP). Excitations from core orbitals or to unoccupied orbitals have been suggested to account for dynamic correlation and orbital relaxation effects^{9,10} and, hence, do not comprise additional spin-coupling mechanisms.

Another modification of the two electrons in two localized orbitals scheme presented above is the addition of a third electron resulting in a mixed valence compound such as the stretched H_2^- molecule. Delocalization or “resonance” of the excess electron between the magnetic sites A and B stabilizes the system and occurs favorably when the local spins \mathbf{S}_A and \mathbf{S}_B are aligned in parallel. This *double exchange* effect may then give effective ferromagnetic contributions to the spin coupling in mixed valence TM complexes.^{11,12}

The method most widely used today to predict J coupling constants for polynuclear complexes is the broken-symmetry approach proposed by Noodleman.^{13,12} In this single configuration description, the solution for the low-spin state (the BS state, corresponding to $M_S = |S_A - S_B|$ in the above example with two magnetic centers) has space and spin symmetries broken. Such state is not a spin eigenstate but a superposition of spin states weighted by Clebsch–Gordan coefficients. A value for J can be estimated¹⁴ by using spin-projection techniques and by also computing the highest spin (HS) state, which usually is well described by a single configuration:

$$J = -\frac{E_{\text{HS}} - E_{\text{BS}}}{\langle \hat{S}^2 \rangle_{\text{HS}} - \langle \hat{S}^2 \rangle_{\text{BS}}} \quad (4)$$

where E is the state energy and $\langle \hat{S}^2 \rangle$ is the expectation value of the total spin operator. The success of the broken-symmetry approach can be traced to appropriate descriptions of direct exchange, superexchange, and LSP mechanisms discussed above.⁸ However, its accuracy obviously depends on the electronic structure method employed for the single configuration calculations, which is often spin-polarized density functional theory. Because eigenfunctions for the lower spin states are not obtained explicitly, the broken-symmetry approach is not suited to study state specific properties. Nevertheless, mapping and spin-projection techniques may also be applied to estimate \mathbf{g} tensors and hyperfine coupling constants¹⁵ and to optimize geometries¹⁶ approximately. Along the same line, an extended broken-symmetry approach has been introduced recently that allows the calculation of energy derivatives for homovalent binuclear complexes.¹⁷

From the VB discussion in the previous paragraphs, it seems evident to employ configuration interaction (CI) of Slater determinants to compute wave functions for low-spin eigenstates. All spin-coupling mechanisms and electronic effects cited above can be naturally accounted for if an appropriate configuration space is used. However, the exponential scaling of the size of the CI space puts serious limitations on the range of TM complexes and properties that can be calculated with CI. For instance, the configurational space generated in full excitation level for about 18 unpaired electrons already exceeds the capacity of modern

CI code implementations and computer hardware. At this point, some heroic CI computations on low-spin states of binuclear TM complexes by Malrieu and collaborators should be mentioned.^{18,9} Their dedicated difference CI method has been used to compute energy differences between spin multiplets in very good agreement with experimental data. Together with perturbative analysis, this CI method has also been used to identify contributions to spin coupling.^{9,10} Even so, the dedicated difference CI also suffers from an exponential scaling of the CI space and thus is limited to binuclear complexes with a small number of unpaired electrons.

In this paper approximate levels of CI selection are proposed in trying to find short CI expansions that still capture the essential physics of spin coupling for the low-spin eigenstates. Determinants are built with localized molecular orbitals. But instead of specifying a given level of excitation from a single reference as in canonical CI, the configurational space is built by completing the spin manifold for neutral (or covalent), ionic, and ligand-to-metal charge-transfer VB-like structures. It is important to note that all approximations proposed here concern only the selection of configurations that enter in the CI. Thus, all the conclusions obtained should be equally valid irrespective of the method, semiempirical or *ab initio*, used to calculate the molecular integrals and configuration energies. A semiempirical Hamiltonian was employed here because future applications of the proposed approximations will use a hybrid quantum/classical potential based on semiempirical methods. Tests are performed in several simple systems so that full CI calculations can be carried out as references. Details of the computational methods are given in the next section. The results show that single ionic excitations between magnetic sites are enough to obtain an accurate superexchange contribution. Further reduction in the size of the CI space is possible but restricts the description to ground spin ladders. For iron–sulfur clusters, spin coupling can be correctly described by rather small CI expansions, paving the way for simulation studies of magnetic and electronic properties of these prosthetic groups in the condensed phase.

2. Computational Methods

Test calculations were performed on simple spin-coupled molecular systems. Two homonuclear diatomics, N_2 and Cr_2 , two bridged triatomics, N_2F^- and Fe_2S^{4+} , and the ring cluster $Fe_2S_2^{2+}$, were studied. Dinitrogen bond distance was set to 4.5 bohr (~ 2.86 Å), and the dichromium bond distance was set to 4.4 bohr (~ 2.33 Å). At such separations, covalent bonding is not significant, and energy splittings between the total spin eigenstates have magnitudes similar to those observed in polynuclear TM complexes. The equilibrium bond lengths for dinitrogen and dichromium are ~ 1.11 and ~ 1.68 Å, respectively. Each atom in the stretched diatomic molecule plays the role of an open-shell metal center or magnetic site. The unpaired electrons are weakly interacting, in a suitable model to the direct exchange and through-space superexchange mechanisms. Yet, dinitrogen is simple enough to allow complete expansions of the electronic wave function as well as several levels of CI selection. Neutral, dipositive, and mononegative total molecular charges were assigned for

dinitrogen as models of magnetic compounds with half-full open shell, less than half-full, and mixed valence, respectively. Triatomic molecules composed of two magnetic centers separated by a diamagnetic ligand are the simplest systems to probe the effect of the proposed approximations on interactions via the ligand spin polarization mechanism. Since bridge ligands found in TM complexes are usually diamagnetic anions, stretched dinitrogen was bridged with fluoride in an angular geometry with C_{2v} symmetry, $\angle = 75^\circ$, $d(N-F) = 1.80$ Å, and $d(N-N) = 2.19$ Å. In the TM compound Fe_2S^{4+} , two Fe(III) are bridged by a sulfide ligand. A symmetric linear geometry was adopted with $d(Fe-S) = 1.271$ Å. The binuclear iron–sulfur cluster $Fe_2S_2^{1+/2}$ is the prosthetic group found in many electron-transfer proteins, such as ferredoxin. Each iron is also attached to the protein by two cysteine sulfur atoms, with a total tetrahedral coordination. By contrast, the bare $Fe_2S_2^{2+}$ cluster studied here, a D_{2h} geometry was used,¹⁹ with $d(Fe-Fe) = 2.543$ and $d(Fe-S) = 2.251$ Å. The z axis contains the two magnetic sites in all molecules studied.

Calculations were carried out with a semiempirical neglect of diatomic differential overlap (NDDO) Hamiltonian.^{20,21} A slightly modified version of the MOPAC2000^{22,23} code that allowed CI calculations using localized molecular orbitals was employed. Standard AM1 parameters were used for nitrogen and fluoride²⁴ and modified neglect of differential overlap (MNDO)-d parameters were used for sulfur.²⁵ MNDO-d parameters were not available for chromium and iron, so a quick parametrization had to be done. See details and the parameter values in the Supporting Information. Molecular orbitals (MOs) were obtained from high-spin restricted open-shell Hartree–Fock (ROHF) calculations and were localized using an equivalent Pipek–Mezey procedure.²⁶ Although MOPAC does not work with symmetry-adapted basis, all resulting wave functions were checked for the correct space and spin symmetries. Active spaces defined for the CASCI (full CI on the given active space)²⁷ calculations contained all open-shell MOs as well as outer valence unoccupied and double-occupied MOs in N_2^{+2} and N_2^- , respectively. All unpaired electrons were included in the active spaces. Full details of the active spaces used are given for each tested molecule in the Results and Discussion Section. Approximate CI expansions were based on the VB arguments presented in the Introduction. Hence, instead of specifying a given level of excitation from the ROHF solution, the selected CI expansions included all determinants needed to complete the spin manifold for a given level of approximation for the mechanisms of effective spin-coupling discussed. Only $M_S = 0$ (or $M_S = 0.5$, for N_2^-) determinants were used in the selected CI expansions.

For the larger active spaces, CASCI calculations were not feasible for the low-spin states (singlet and triplet). MOPAC generates and diagonalizes the CI matrix (or secular determinant) explicitly, and the code could not be compiled to use more than 2 GB of memory. Thus, the size of the CI expansions were limited to about 9000 configurations, which is less than the number of configurations necessary to expand the singlet and triplet states for the molecules formed by Cr and Fe. All the CASCI calculations were done with the

Table 1. Relative Energies (eV) and Number of Configurations (size) Included in the Wavefunction Expansions for Electronic Eigenstates of Dinitrogen in Neutral, Dipositive, and Negative Total Molecular Charge

		N_2^0			
CI expansion	size	$1\Sigma_g$	$3\Sigma_u$	$5\Sigma_g$	$7\Sigma_u$
CASCI	400	0.0000	0.0281	0.0876	0.1872
neu + single ion	80	0.0000	0.0278	0.0868	0.1864
neu + p_x, p_y, p_z ion	56	0.0000	0.0278	0.0868	0.1864
neu + p_z ion	32	0.0000	0.0262	0.0822	0.1778
neu + p_x, p_y ion	44	0.0000	0.0014	0.0043	0.0087
		N_2^{2+}			
CI expansion	size	$1\Sigma_u$	$3\Sigma_g$	$5\Sigma_u$	
CASCI	225	0.0000	0.0515	0.1953	
neu + single ion	162	0.0000	0.0514	0.1939	
neu + unpair, p_z ion	114	0.0000	0.0514	0.1939	
neu p_z + unpair, p_z ion	72	0.0000	0.0514	0.1939	
neu + p_z ion	78	0.0000	0.0567	0.2028	
neu + unpair ion	90	0.0000	-0.0029	-0.0087	
		N_2^-			
CI expansion	size	$2\Sigma_u$	$4\Sigma_g$	$6\Sigma_u$	
CASCI	300	0.3125	0.1458	0.0000	
neu + single ion	240	0.3126	0.1458	0.0000	
neu p_z + p_z ion	44	0.3130	0.1460	0.0000	
neu	60	0.3298	0.1568	0.0000	
neu p_z	20	0.3298	0.1568	0.0000	

semiempirical NDDO Hamiltonian. The CASSCF method²⁷ within the MOLCAS 6.2 program system²⁸ was used to compute a reference value for the $Fe_2S_2^{2+}$ cluster. This calculation was conducted with basis symmetry adapted to the D_{2h} point group, using the ANO-RCC²⁹ set with quadruple- ζ contraction (e.g., 7s6p4d3f2g for iron).

3. Results and Discussion

Results of several approximate levels of CI selection on the electronic structure of simple molecules are presented in this section. For the diatomic systems and the linear Fe_2S^{4+} , all spin ladders shown are Σ states. For NFN^- , the lowest energy spin states are alternating A_1 and B_2 states, and for the ring Fe_2S_2 , the spin ladder shown has alternating A_g and B_{1u} states. For example, the correct energy ordering for the total spin eigenstates of neutral N_2 is $1\Sigma < 3\Sigma < 5\Sigma < 7\Sigma$.

3.1. Neutral N_2 . For neutral N_2 , the following configuration is obtained after localizing the high-spin ROHF MOs: $[(core)2s^A 2s^A 2s^B 2s^B 2p_z^A 2p_z^B 2p_x^A 2p_x^B 2p_y^A 2p_y^B]$, where the over bar assigns spin down and the superscripts A and B are used to label each nitrogen atom. Localized MOs have large contributions by only one atomic function which is then used as a label. The six unpaired electrons in the six 2p MOs are responsible for the spin coupling and form the active space for generation of configurations used in the wave function expansion. Because of localization, the MOs will have a $C_{\infty v}$ symmetry, which is lower than the nuclear point group.

The relative energies obtained for the lowest energy spin eigenstates are shown in Table 1. The CASCI has a total of 400 configurations with $M_S = 0$. There are 20 unpaired neutral configurations, i.e., with 1 electron in each of the 6

active MO. The septet wave function is composed only by these 20 configurations, with equal CI weights. The largest CI weights (~ 0.24 in the singlet state) in the expansions for all other spin states come from two configurations, $|p_z^A p_z^B p_x^A p_x^B p_y^A p_y^B|$ (only the active space is represented on this and the following determinant configurations) and the respective A to B spin inversion. These two configurations correspond to a $4S$ high-spin state on each N atom. The second largest contributions come from the other 18 unpaired neutral configurations, such as $|p_z^A p_z^B p_x^A p_x^B p_y^A p_y^B|$, which corresponds to combinations of atomic excited states or non-Hund states.³⁰ Ionic configurations have rather smaller contributions (CI weight ≤ 0.03 in the singlet). The next-lying excited state above the septet shown in Table 1 is at least 2 eV higher in energy.

Judgement from the weights in the CASCI expansion would suggest that only the 20 unpaired neutral configurations could be used in the wave function expansion for all spin eigenstates. However, this approximation results in a flat spin ladder, with the same energy for all states. As described in the Introduction Section, neutral configurations are not able to account for the effective antiferromagnetic interactions between the open shells. The ladder is flat because MOs are strictly localized so that the direct exchange (K_{ab}) ferromagnetic contribution is very small, actually null in the precision used. The first reasonable level of approximation, named neu + single ion in Table 1, is an expansion containing 20 neutral configurations plus all the 60 symmetry-allowed “metal-to-metal” (or nitrogen-to-nitrogen) ionic single excitations that can be constructed from the set of neutral configurations, e.g., $|p_z^B p_z^A p_x^B p_x^A p_y^B p_y^A|$. The energy values obtained with this expansion are within 0.001 eV of the CASCI reference, and the number of configurations used is five-fold smaller. Since localized MOs are used, excitations between MOs that belong to the same irrep of $C_{\infty v}$ are symmetry allowed. A second approximation can be made by including neutral and single ionic excitations only between localized MOs composed by the same atomic functions (neu + p_x, p_y, p_z ion, Table 1). This results in identical energies showing that symmetry-allowed “crossed” ionic excitations (e.g., $p_x^B \rightarrow p_y^A$) do not interact with the wave function for the low-lying states of neutral N_2 . An expansion including neutral and the 12 single ionic excitations between the $2p_z$ MOs (neu + p_z ion) results in energies within 0.01 eV of the CASCI reference. This suggests a third level of approximation in which the only ionic excitations included are those between MOs composed of atomic functions with large overlap (the z axis is the intermolecular axis). As a counter example, an expansion including neutral and ionic excitations between MOs composed of atomic functions with small overlap (neu + p_x, p_y ion) results in almost no antiferromagnetic contributions and a spin ladder in large disagreement with the CASCI reference. It should be noted that, by progressively removing from the CI space the excitations between $2p_x$ and $2p_y$ MOs (as in neu + p_x, p_y, p_z ion and in neu + p_z ion), spin ladders of higher energy and different space symmetry will not be correctly described. This is not a problem for neutral N_2 because the next-lying state above the 7Σ state is much higher in energy, but it might

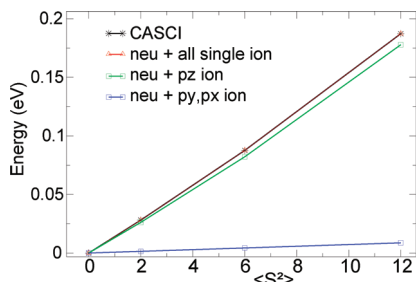


Figure 1. Spin ladders for the lowest energy total spin eigenstates of N_2^0 calculated with different wave function expansions. See text for details.

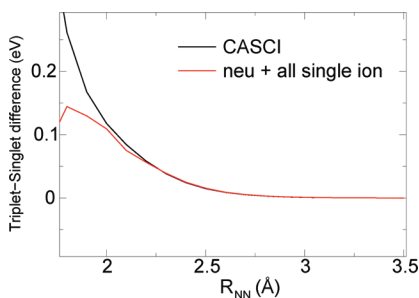


Figure 2. Triplet-singlet energy gap for varying N_2^0 bond distances.

introduce errors when the ground spin ladder is near degenerate to other ladders.

Double ionic ($N^{2-}-N^{2+}$), triple ionic ($N^{3-}-N^{3+}$), and internal paired neutral configurations, e.g. $|p_z^A p_z^B p_x^A p_x^B\rangle$, which also corresponds to non-Hund atomic states, have very small contributions and can be safely neglected. Removing the two neutral configurations corresponding to the 4S high-spin state on each N atom from the expansion neu + single ion or using only these two neutral configurations plus all single ionic ones results in an incomplete spin manifold and, consequently, bogus spin ladders.

Linear spin ladders, i.e., ladders that follow a regular Landé splitting, are obtained within the CASCI, and the levels of approximation suggested above are shown in Figure 1. The CASCI ladder and the expansion named neu + single ion have both correlation coefficients to a straight line of 0.9994 and a F variance quality of 1662. The expansion neu + p_z ion has a correlation of 0.9992 and a F variance quality of 1187. In conclusion, the CI expansion neu + single ion captures the essential physics of exchange interactions for the ground spin ladder (Table 1) as well as for higher energy ladders (not shown) of the stretched dinitrogen molecule.

To test the limits of the proposed configuration selection, the singlet-triplet energy gap was calculated with varying bond distances. Figure 2 shows that the expansion neu + single ion results in energy gaps in very good agreement with the CASCI wave function down to bond distances of ~ 2.0 Å. Below this distance, the interaction between the unpaired electrons is strong, and covalent bonding becomes appreciable. The system is not only spin coupled, and the proposed approximate CI selections do not apply.

3.2. N_2^{2+} . For N_2^{2+} , the configuration obtained after localizing the high-spin ROHF MOs is equivalent to the neutral N_2 configuration (see above) but with two previous

highest occupied molecular orbitals (HOMOs) now unoccupied. The relative energies obtained for the lowest energy spin eigenstates are shown in Table 1. The expansion neu + single ion results in energy values in excellent agreement (within 0.002 eV) with the CASCI reference. For less than half-filled open shells, there are ionic configurations which still have all electrons unpaired. There are 36 of such unpaired ionic configurations for N_2^{2+} . An expansion including all neutral configurations, unpaired ionic and single ionic excitations between the $2p_z$ MOs (neu + unpair, p_z ion) result in energies identical to the neu + single ion expansion. Single ionic excitations between MOs composed by atomic functions with small overlap (e.g., $p_y^B \rightarrow p_y^A$) and crossed single excitations do not interact with the wave function for the low-lying states of N_2^{2+} . A selection of the neutral configurations included in the expansions is possible for the open-shell systems without exactly half-full shells, i.e., more or less than half-filled and mixed valence. An expansion including only neutral configurations with one electron in each $2p_z$ MOs, unpaired ionic and single ionic excitations between the $2p_z$ MOs (neu p_z + unpair, p_z ion) also result in energies identical to the neu + single ion expansion. An expansion including all neutral configurations and the 24 single ionic excitations between the $2p_z$ MOs (neu + p_z ion) results in energies within 0.01 eV of the CASCI reference. But, contrary to the equivalent neu + p_z ion expansion for the neutral N_2 , an excess antiferromagnetic character is observed. This is a consequence of neglecting the ferromagnetic contribution of unpaired ionic configurations, easily seen in the results for the neu + unpair ion expansion in Table 1. Thus, not all metal-to-metal ionic excitations give an antiferromagnetic contribution to spin coupling, but only those that alter the number of unpaired electrons.

Considering a particle-hole symmetry, an equivalent behavior would be observed for the more than half-filled case. For example, in N_2^{2-} , ionic configurations without an empty MO give ferromagnetic contributions, equivalent to the ionic unpaired configurations in the less than half-filled case.

3.3. N_2^- . For N_2^- , the localized high-spin ROHF MOs used in the CI expansions were obtained for the neutral dinitrogen to avoid an artificial polarization of the occupied MOs and thereof biased CI results. Similar results were obtained if a fractional occupation of the MOs was allowed in the ROHF solution. The relative energies obtained for the lowest energy spin eigenstates are shown in Table 1. Delocalization of the excess electron stabilizes the “neutral” configurations resulting in a ferromagnetic CASCI spin ladder. This is the double-exchange effect.¹¹ Antiferromagnetic contributions by the superexchange mechanism are an order of magnitude smaller. Thus, an expansion including only neutral configurations (neu, Table 1) accounts for the double-exchange effect and results in energies within 0.02 eV of the CASCI reference. In fact, an expansion (neu p_z) in which the excess electron occupies only the $2p_z$ orbitals has identical results. However, by removing from the CI space configurations in which the excess electron occupies the $2p_x$ and $2p_y$ MOs, spin ladders

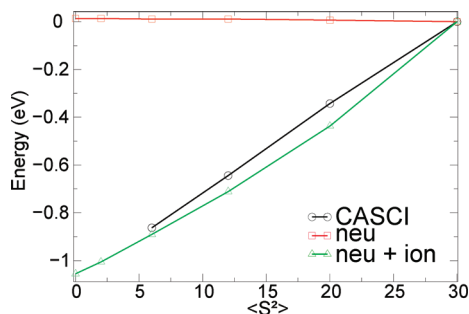


Figure 3. Spin ladders for the lowest energy total spin eigenstates of Cr_2 at 4.4 bohr separation calculated with different wave function expansions.

with higher energy and different space symmetry will not be correctly described, as observed for similar CI selections in $\text{N}_2^{0/+2}$.

The antiferromagnetic contribution can be retrieved in an expansion including all symmetry-allowed single ionic excitations (neu + single ion) resulting in energies within 0.0001 eV of the CASCI reference. The expansion including the interacting neutral and the 24 single ionic excitations between the $2p_z$ MOs (neu $p_z + p_z$ ion) contains five-fold less configurations than the CASCI and results in energies within 0.001 eV of this reference.

3.4. Cr_2 . For the stretched dichromium molecule, covalent bonding between the 3d orbitals is not significant. However, there is still a σ bond formed mostly between the diffuse 4s chromium orbitals.³¹ The correct energy ordering for the total spin states should have the antiferromagnetic singlet as the ground state and the ferromagnetic undecaplet as the highest energy state of the ground spin ladder.

The canonical high-spin ROHF solution has 10 singly occupied MOs formed by antisymmetric and symmetric combinations of the atomic 3d functions. The HOMO-1 and HOMO are formed, respectively, by antisymmetric and symmetric combinations of the 4s functions. After full orbital localization, each Cr atom contains six electrons and a configuration corresponding to a ^7S atomic state. The active space was composed of the 12 electrons in 10 MOs formed by 3d functions and the 2 MOs formed by 4s functions. All the configurations used in the expansions were formed out of the two possible combinations of the localized 4s orbitals consistent with a σ bonding MO. The CASCI solution was only computed down to the quintet state. The secular determinants necessary to obtain states with $S < 2$ were too large and could not be built due to memory limitations (see Computational Methods Section).

Figure 3 shows spin ladders calculated for Cr_2 under different CI selections. The undecaplet state was chosen as zero of energy. An expansion including only unpaired neutral configurations (252 configurations in total), e.g., $[\text{core}]d_z^A d_z^B d_x^A d_x^B d_y^A d_y^B d_{xz}^A d_{xz}^B d_{yz}^A d_{yz}^B$, yields an incorrect spin ladder with a high-spin ground state, as expected from the direct-exchange contribution to spin coupling. The expansion neu + single ion, including all 252 unpaired neutral configurations plus 1260 ionic configurations, results in fair agreement with the CASCI result (within 0.1 eV). Ionic configurations were built from the set of unpaired

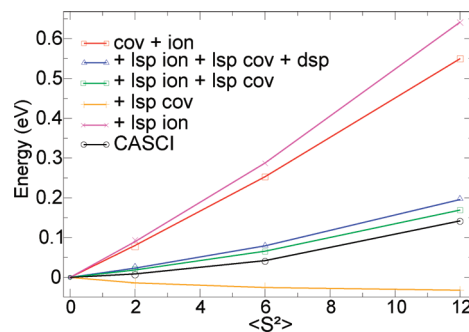


Figure 4. Spin ladders for the lowest energy total spin eigenstates of NFN^- in angular geometry calculated with different wave function expansions.

neutral configurations by metal-to-metal single excitations between MOs belonging to the same irreps of the $C_{\infty v}$ group. It should be noted that the neu + single ion CI expansion contains only 1512 configurations, instead of the 63 504 configurations that would be necessary to expand the singlet state in the CASCI wave function. Energies within 0.01 eV of the neu + single ion expansion are obtained by a smaller expansion with 952 configurations that does not contain the crossed ionic excitations between MOs belonging to the same $C_{\infty v}$ irrep but formed by different atomic functions, e.g., $d_{xz}^A \rightarrow d_{yz}^B$.

The spin ladders obtained with CASCI and neu + single ion approximation have, respectively, correlation coefficients to a straight line of 0.9998 and 0.996 and a F variance quality of 4032 and 465. The approximations proposed for the model stretched dinitrogen are equally valid for the stretched dichromium and result in a reduction of at least two orders of magnitude in the size of the CI space.

3.5. NFN^- in C_{2v} Symmetry. On the following sections, the proposed approximations are tested on compounds containing diamagnetic bridges. Localized MOs were obtained for angular NFN^- from a high-spin ROHF solution. Each nitrogen has a double-occupied 2s-like shell and 3 unpaired electrons in orbitals composed by the 2p functions. Fluoride has 4 double-occupied orbitals composed by 2s and 2p functions. All 9 MOs formed by p functions and 12 electrons are included in the active space.

Figure 4 shows spin ladders calculated under different CI selections. The singlet was chosen as zero of energy. An expansion including only unpaired neutral and ionic single excitations between the magnetic (nitrogen) centers with the bridge (fluoride) MOs left double occupied (neu + single ion, 80 configurations in total) yields a largely antiferromagnetic ladder, in large disagreement with the CASCI reference. The ligand spin polarization has to be included for a qualitatively correct description of the NFN^- wave function.

LSP configurations are obtained by LMCT *single* excitations built from the set of neutral and ionic determinants. An expansion including the neu + single ion set and all configurations generated from the (20) neutral determinants by LMCT single excitations (+ lsp neu, 280 configurations in total) results in an overestimation of the ferromagnetic interactions. On the other hand, a similar expansion (+ lsp ion, 406 configurations in total) but with LSP configurations

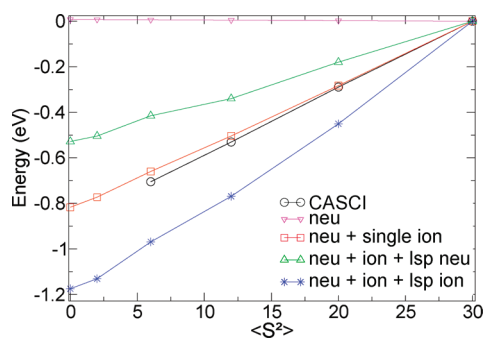


Figure 5. Spin ladders for the lowest energy total spin eigenstates of linear Fe_2S^{4+} calculated with different wave function expansions.

generated by single excitation from the (60) ionic determinants results in an overestimation of the antiferromagnetic contributions. A proper balance is obtained by an expansion including the neu + single ion set plus both neutral and ionic ligand spin polarization (+ lsp ion + lsp neu), resulting in energies within 0.03 eV of the CASCI reference. Double LMCT excitations can also be constructed from the neutral and ionic configurations set by either exciting the same bridge orbital or two different ones. The resulting contributions are anti- and ferromagnetic but with much smaller magnitude (~ 0.01 eV), as found for other TM bridged systems.^{8,7}

In other words, the set of unpaired neutral and single ionic configurations might be considered a zero-order reference set. LSP excitations built out of this multireference set result in anti- and ferromagnetic contributions to spin coupling if the excitation originates from an ionic or a neutral configuration, respectively. For NFN^- , the LSP contributions are larger than the through-space superexchange contributions and have to be included for a qualitatively correct description of the spin coupling. This is not generally true, as shown below for the TM compounds. DSP contributions are relatively small and can be removed from the CI space without affecting the results significantly.

3.6. Linear Fe_2S^{4+} . Localized MOs were obtained for linear Fe_2S^{4+} from a high-spin ROHF solution. Each metal center has a half-filled valence shell with 5 unpaired electrons in 5 orbitals composed by 3d functions, corresponding to an atomic ^6S state. The sulfur bridge has four double-occupied 2s- and 2p-like orbitals. The 3 outer-valence bridge MOs and the 10 MOs formed by iron 3d functions were included in the active space, with the respective 16 electrons.

A CASCI solution with such a large active space is not feasible within the memory limitations found here (see Computational Methods Section). Instead, Figure 5 shows a CASCI result obtained with only 10 electrons in the 10 MOs formed by iron 3d functions. The undecaplet state was chosen as zero of energy. The neu expansion includes only unpaired neutral configurations with double-occupied ligand MOs (252 configurations in total, Figure 5) and results in small ferromagnetic coupling, as observed above for the chromium dimer. The neu + single ion expansion includes the ionic configurations (1512 configurations in total) and results in very good agreement with the 10 electron in 10 orbitals CASCI. The effect of neutral LSP configurations (neu + ion + lsp neu) is ferromagnetic, and the ionic LSP (neu + ion

Table 2. Relative Energies (eV) for $\text{Fe}_2\text{S}_2^{2+}$ Lowest Energy Spin Eigenstates Calculated with Two Different CI Expansions

$\langle \hat{S}^2 \rangle$	neu + single ion	CASSCF
0	0.000	0.000
2	0.041	0.046
6	0.135	0.136
12	0.283	0.267
20	0.475	0.427
30	0.717	0.679

+ lsp ion) is antiferromagnetic. Neutral and ionic LSP configurations are obtained by LMCT *single* excitations built from the set of neutral (neu) and ionic (ion) determinants, respectively. However, contrary to the NFN^- example above, through-space superexchange dominates, and the ligand spin polarization is relatively smaller in Fe_2S^{4+} . For example, the ladder obtained with the expansion neu + ion + lsp neu is antiferromagnetic. In fact, inclusion of both neutral and ionic LSP configurations practically cancels out the polarization effect and results in a spin ladder very close to the neu + single ion expansion. Even if LMCT excitations are not explicitly included in the CI space, the effect of bridges and ligands is at least partially included when MOs are generated and when energies of the zero-order multireference configurations are calculated. Results similar to those shown in Figure 5 for the neu + single ion expansion are obtained by removing the crossed ionic excitations, as observed above for stretched N_2 and Cr_2 , leading an expansion with only 952 configurations.

3.7. $\text{Fe}_2\text{S}_2^{2+}$ Ring. The final example is the iron–sulfur cluster $\text{Fe}_2\text{S}_2^{2+}$. Localized MOs obtained from a high-spin ROHF solution show a half-filled valence 3d shell with 5 unpaired electrons in each iron center. Each sulfur bridge has 3 outer-valence double-occupied localized MOs composed by 2p functions. An active space containing all 16 valence MOs and the respective 22 electrons is only feasible using modern direct CI procedures. An ab initio CASSCF computation using such large active space is taken as reference in Table 2. The CASSCF singlet wave function is expanded in almost two million determinants in comparison to the approximate and much shorter expansion neu + single ion that includes only 1512 determinants corresponding to the unpaired neutral and ionic single excited states. The agreement between the CASSCF and the selected neu + single ion expansion is very good (within 0.05 eV) and suggests that this level of approximation captures the essential physics of spin coupling in transition-metal complexes. In fact, energies within 0.003 eV of the neu + single ion expansion were obtained with an even smaller expansion containing 952 configurations, by removing the crossed ionic excitations.

4. Conclusions

Approximate configuration interaction expansions were introduced for the calculation of wave functions with correct spin and space symmetries of weakly coupled transition-metal compounds with many open shells. The selection of configurations included in the CI space was based on physical

arguments for the mechanisms of spin coupling, namely direct exchange, superexchange, double exchange, and ligand spin polarization. In the spirit of valence-bond calculations, localized (molecular) orbitals were used in the construction of Slater determinants. But, instead of specifying a level of excitation as in the normal CI terminology, the expansions included all determinants needed to complete the spin manifold compatible with the exchange mechanisms depicted in the Introduction Section.

A zero-order multireference set was identified as the set of neutral and single ionic configurations. The neutral set accounts for the direct-exchange ferromagnetic mechanism and corresponds to configurations with an equivalent number of unpaired electrons in each magnetic site (excluding the excess electron in mixed valence systems). The ionic set is built by symmetry-allowed metal-to-metal single excitations from the neutral set that alter the total number of unpaired electrons. For all the spin-coupled compounds tested here and, we believe, for any spin-coupled system, single ionic excitations are enough to account for the through-space superexchange antiferromagnetic mechanism.

Symmetry-allowed excitations involve molecular orbitals that belong to the same irrep of the *localized* MOs point group. The contribution of symmetry-allowed crossed ionic excitations, i.e., excitations between MOs formed mainly by different atomic functions, was very small or null for the ground spin ladder in all molecules studied. For other systems, this result will depend on the localization method employed and on whether the localized MOs resemble pure atomic orbitals or combinations thereof. Even smaller expansions are possible by selectively removing from the CI space other ionic configurations or neutral configurations for the more or less than half-filled and mixed valence systems that have very small or null CI weights in the expansions of the low-lying spin states. For instance, removing excitations between MOs formed by atomic functions with a small overlap in N₂ resulted in energies close to those obtained with the full zero-order set for the ground spin ladder. However, spin ladders of higher energy and different space symmetry might not be correctly described by CI spaces smaller than the zero-order multireference set.

Ligand-to-metal charge-transfer configurations constructed from the zero-order reference set account for ligand spin polarization and double spin polarization. Single LMCT out of the neutral set always give a ferromagnetic contribution. On the other hand, single LMCT out of the ionic set always give an antiferromagnetic contribution, sometimes called “through-bond” superexchange. The LSP configurations should be included in the CI space whenever this contribution is comparable in magnitude to through-space direct and superexchange. For the iron–sulfur compounds studied here, the LSP contribution is small and approximately cancels out when both ionic and neutral single LMCT excitations are included. This is not a general result,^{9,10,30} but it is a valuable one in reducing the size of the CI expansions. A related argument is valid for the mixed valence system tested. The double-exchange effect in N₂⁻ is much larger

than the superexchange so that ionic configurations can be excluded from the CI space without affecting the energy splittings significantly.

Comparisons with experimental *J* coupling constants are not given here. Such comparisons would not be fair at this stage because the calculations presented do not include the effect of dynamic correlation, which is essential for quantitative results.^{27,9} Dynamic correlation can be added on top of the zero-order set by either multireference CI or perturbative corrections.²⁷ If a semiempirical method is employed, then correlation can be implicitly included in the parametrization of electron-repulsion integrals.

The proposed approximations result in much shorter CI expansions. For example, the CASSCF result obtained with 2×10^6 configurations for Fe₂S₂²⁺ is reproduced with about 10³ configurations. However, the exponential scaling of the CI space size is not entirely ameliorated. Polynuclear compounds with a larger number of magnetic centers and unpaired electrons will still require large configurational spaces that may exceed the available computational resources even if including only neutral and single ionic excitations between neighboring sites. Nevertheless, identifying the spin-coupling mechanisms with valence-bond structures and including controlled approximations in the CI expansion may open the way to treat these more challenging systems.

Acknowledgment. Jeppe Olsen (University of Århus) is acknowledged for helpful discussions. G.M.A. acknowledges funding from FAPESP, projects 07/52772-6 and 07/59345-6.

Supporting Information Available: Semiempirical parameters and the procedure used for their calibration. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Blondin, G.; Girerd, J.-J. *Chem. Rev.* **1990**, *90*, 1359–1376.
- (2) Heisenberg, W. *Z. Phys.* **1928**, *49*, 619–636.
- (3) Dirac, P. *Proc. R. Soc. London, Ser. A* **1929**, *123*, 714–733.
- (4) Anderson, P. W. *Phys. Rev.* **1950**, *79*, 350–356.
- (5) Anderson, P. W. *Phys. Rev.* **1959**, *115*, 2–13.
- (6) Hay, P. J.; Thibault, J. C.; Hoffmann, R. *J. Am. Chem. Soc.* **1975**, *97*, 4884–4899.
- (7) de Loth, P.; Cassoux, P.; Daudey, J. P.; Malrieu, J. P. *J. Am. Chem. Soc.* **1981**, *103*, 4007–4016.
- (8) Noodleman, L.; Davidson, E. R. *Chem. Phys.* **1986**, *109*, 131–143.
- (9) Calzado, C. J.; Cabrero, J.; Malrieu, J. P.; Caballol, R. *J. Chem. Phys.* **2002**, *116*, 2728–2747.
- (10) Calzado, C. J.; Angeli, C.; Taratiel, D.; Caballol, R.; Malrieu, J.-P. *J. Chem. Phys.* **2009**, *131*, 044327.
- (11) Zener, C. *Phys. Rev.* **1951**, *82*, 403–405.
- (12) Noodleman, L.; Han, W.-G. *J. Biol. Inorg. Chem.* **2006**, *11*, 674–694.
- (13) Noodleman, L. *J. Chem. Phys.* **1981**, *74*, 5737–5743.

- (14) Soda, T.; Kitagawa, Y.; Onishi, T.; Takano, Y.; Shigeta, Y.; Nagao, H.; Yoshioka, Y.; Yamaguchi, K. *Chem. Phys. Lett.* **2000**, *319*, 223–230.
- (15) Sinnecker, S.; Neese, F.; Noodleman, L.; Lubitz, W. *J. Am. Chem. Soc.* **2004**, *126*, 2613–2622.
- (16) Li, J.; Noodleman, L. In *Spectroscopic Methods in Bioinorganic Chemistry*; Solomon, E. I., Hodgson, K. O., Eds.; ACS Symposium Series: Washington, DC, 1998; Vol. 692.
- (17) Nair, N. N.; Schreiner, E.; Pollet, R.; Staemmler, V.; Marx, D. *J. Chem. Theory Comput.* **2008**, *4*, 1174–1188.
- (18) Cabrero, J.; Amor, N. B.; de Graaf, C.; Illas, F.; Caballol, R. *J. Phys. Chem. A* **2000**, *104*, 9983–9989.
- (19) Hubner, O.; Sauer, J. *J. Chem. Phys.* **2002**, *116*, 617–628.
- (20) Dewar, M. J.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.
- (21) Thiel, W. Semiempirical Methods. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; John von Neumann Institute for Computing: Jülich, Germany, 2001; Vol. 3, pp 1–24.
- (22) Stewart, J. J. P. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–45.
- (23) Stewart, J. J. P. *MOPAC 2000*; Fujitsu Limited, Tokyo, Japan, 1999.
- (24) Dewar, M. J.; Zoebisch, E. G.; Healy, H. F.; Stewart, J. P. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (25) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100*, 616–626.
- (26) Pipek, J.; Mezey, P. *J. Chem. Phys.* **1989**, *90*, 4916–4926.
- (27) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*, 1st ed.; Wiley: New York, 2000; pp 523–645.
- (28) Karlstrom, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.
- (29) Roos, B. O.; Lindh, R.; Malmqvist, P.-A.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2005**, *109*, 6575–6579.
- (30) Bastardis, R.; Guihéry, N.; de Graaf, C. *J. Chem. Phys.* **2008**, *129*, 104102.
- (31) Goodgame, M. M.; Goddard, W. A., III *Phys. Rev. Lett.* **1985**, *54*, 661–664.

CT1001279

JCTC

Journal of Chemical Theory and Computation

A System-Dependent Density-Based Dispersion Correction

Stephan N. Steinmann and Clemence Corminboeuf*

Laboratory for Computational Molecular Design, Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

Received March 18, 2010

Abstract: Density functional approximations fail to provide a consistent description of weak molecular interactions arising from small electron density overlaps. A simple remedy to correct for the missing interactions is to add *a posteriori* an attractive energy term summed over all atom pairs in the system. The density-dependent energy correction, presented herein, is applicable to all elements of the periodic table and is easily combined with any electronic structure method, which lacks the accurate treatment of weak interactions. Dispersion coefficients are computed according to Becke and Johnson's exchange-hole dipole moment (XDM) formalism, thereby depending on the chemical environment of an atom (density, oxidation state). The long-range $\sim R^{-6}$ potential is supplemented with higher-order correction terms ($\sim R^{-8}$ and $\sim R^{-10}$) through the universal damping function of Tang and Toennies. A genuine damping factor depending on (iterative) Hirshfeld (overlap) populations, atomic ionization energies, and two adjustable parameters specifically fitted to a given DFT functional is also introduced. The proposed correction, dDXDM, dramatically improves the performance of popular density functionals. The analysis of 30 (dispersion corrected) density functionals on 145 systems reveals that dDXDM largely reduces the errors of the parent functionals for both inter- and intramolecular interactions. With mean absolute deviations (MADs) of 0.74–0.84 kcal mol⁻¹, PBE-dDXDM, PBE0-dDXDM, and B3LYP-dDXDM outperform the computationally more demanding and most recent functionals such as M06-2X and B2PLYP-D (MAD of 1.93 and 1.06 kcal mol⁻¹, respectively).

Introduction

Kohn–Sham density functional theory (DFT)¹ offers a powerful and robust methodology for investigating electronic structures of many-body systems, providing a practical balance of accuracy and computational cost unmatched by other methods. Despite this success, the commonly used semilocal approximations have difficulties in properly describing attractive dispersion interactions that decay with R^{-6} at large intermolecular distances. Even in the short to medium range, most semilocal density functionals fail to give an accurate description of weak interactions.^{2–4}

Accurate treatment of weakly interacting systems is crucial, especially in the field of biomolecules (stacking of

DNA,⁵ protein folding⁶), host–guest chemistry, surface chemistry, and condensed phases of organic molecules. Yet, even seemingly innocuous looking reactions such as alkane isomerization energies and Pople's isodesmic bond separation equations (BSEs),^{7,8} where formal bond types are preserved, suffer from errors at standard DFT levels.^{9–12}

SAPT (DFT)^{13–15} gives highly accurate interaction energies for two or three interacting closed-shell subsystems, but the method is not applicable to intramolecular interactions. Around the energy minimum, dispersion-corrected atom-centered potentials (DCAPs)^{16–22} or specifically fitted density functionals^{23–28} have led to satisfactory results. Nevertheless, both approaches intrinsically lack the ability to recover the long-range $\sim R^{-6}$ attractive form. Conceptually, the simplest remedy is to correct for the missing

* Corresponding author e-mail: clemence.corminboeuf@epfl.ch.

interaction *a posteriori* by adding an attractive energy term summed over all atom pairs in the system. The strategy was originally proposed to improve Hartree–Fock energies (known as HF-D)^{29–32} and was later applied to DFT.^{2–4,33} With parameters for most elements in the periodic table, Grimme’s parametrization³⁴ is the best known DFT-D variant. Since then, there has been considerable interest in finding an optimal parametrization.^{34–52} DFT-D is generally accurate for the treatment of intermolecular interactions, but proper description of weak intramolecular interactions is trickier.^{12,53,54} Specific fitting to a suitable training set⁴⁰ decreases the “intramolecular” error, albeit we have recently shown that the two parametrizations can be unified using a physically motivated damping function called dD10.⁴⁶

Our dD10 correction⁴⁶ is, however, restricted to only a few elements (H, C, N, O) and, like most DFT-D schemes, employs system-independent dispersion coefficients. The present work overcomes these limitations by combining the efficiency of a new damping criterion with the attractiveness of deriving system-dependent dispersion coefficients. Akin to our former correction,⁴⁶ two damping functions are used jointly to treat both intra- and intermolecular weak interactions consistently. System-dependent dispersion coefficients are computed on the basis of the analytical approximation of the Becke and Johnson^{37,55–60} (BJ) exchange-hole-dipole moment (XDM) formalism.^{61,62} Iterative Hirshfeld weights⁶³ are used to partition the dispersion coefficients among the atoms.^{43,64} A genuine and universal damping criterion based on iterative Hirshfeld weights is introduced for the first time. Our approach has the additional advantage of easily incorporating higher order dispersion coefficients absent in, for instance, the related C₆-only scheme of Tkatchenko and Scheffler.⁴⁵ With only two fit parameters, this new dDXDM correction solves difficulties arising from elements positioned in different chemical environments (i.e., selecting a dispersion coefficient^{33–35}) and is easily applicable to every element of the periodic table.

The next sections give details on the implementation and computations. The performance of dDXDM, on test sets featuring both intra- and intermolecular weak interactions, is then compared with the interaction energies of (un)corrected popular functionals (BP86,^{65–67} BLYP,^{65,68} BHHLYP,⁶⁹ B3LYP,^{70,71} PBE,⁷² and PBE0^{73,74}) and established DFT-methods designed to better describe weak interactions (B97-D,³⁴ B2PLYP-D,^{75,76} and M06–2X²⁴).

Theory

The basic form of our correction is the Tang and Toennies (TT) damping function⁷⁷

$$E_{\text{disp,dDXDM}} = - \sum_{i=2}^N \sum_{j=1}^{i-1} \sum_{n=3}^5 f_{2n}(bR_{ij}) \frac{C_{2n}^{ij}}{R_{ij}^{2n}} \quad (1)$$

where N is the number of atoms in the system and b is the TT-damping factor (*vide infra*). The correction is called dDXDM6 if only the first term is retained in the multipole expansion ($n = 3$, corresponding to C₆) and is called dDXDM otherwise ($n = 5$, up to C₁₀). $f_{2n}(bR_{ij})$ represents the “universal

damping functions”⁷⁷ that are specific to each dispersion coefficient and that serve to attenuate the correction at short internuclear distances to account for overlapping densities.

$$f_{2n}(x) = 1 - \exp(-x) \sum_{k=0}^{2n} \frac{x^k}{k!} \quad (2)$$

This coming section describes the procedure employed for the determination of the two nontrivial arguments of eq 1: (i) the dispersion coefficients and (ii) the damping factor b .

i. Dispersion Coefficients and Atomic Partitioning Weights. Dispersion coefficients are computed according to Becke and Johnson’s XDM formalism,^{37,55–60} as efficiently implemented in Q-Chem by Kong and co-workers.^{61,62} The C_6^{ij} , C_8^{ij} , and C_{10}^{ij} coefficients between atoms i and j are, for instance, obtained according to

$$C_6^{ij} = \frac{\alpha_i \alpha_j \langle M_1^2 \rangle_i \langle M_1^2 \rangle_j}{\alpha_j \langle M_1^2 \rangle_i + \alpha_i \langle M_1^2 \rangle_j} \quad (3)$$

$$C_8^{ij} = \frac{3 \alpha_i \alpha_j (\langle M_1^2 \rangle_i \langle M_2^2 \rangle_j + \langle M_2^2 \rangle_i \langle M_1^2 \rangle_j)}{2 (\alpha_j \langle M_1^2 \rangle_i + \alpha_i \langle M_1^2 \rangle_j)} \quad (4)$$

$$C_{10}^{ij} = 2 \frac{\alpha_i \alpha_j (\langle M_1^2 \rangle_i \langle M_3^2 \rangle_j + \langle M_3^2 \rangle_i \langle M_1^2 \rangle_j)}{\alpha_j \langle M_1^2 \rangle_i + \alpha_i \langle M_1^2 \rangle_j} + \frac{21}{5} \frac{\alpha_i \alpha_j \langle M_2^2 \rangle_i \langle M_2^2 \rangle_j}{\alpha_j \langle M_1^2 \rangle_i + \alpha_i \langle M_1^2 \rangle_j} \quad (5)$$

where α_i are atomic polarizabilities and $\langle M_l^2 \rangle$ atomic expectation values of squared multipoles ($l = 1, 2, 3$ for dipoles, quadrupoles, and octupoles, respectively) given by

$$\langle M_l^2 \rangle_i = \sum_{\sigma} \int w_i(r) \rho_{\sigma}(r) [r_i^l - (r_i - d_{X\sigma})^l]^2 d^3r \quad (6)$$

In eq 6, $\rho_{\sigma}(r)$ is the spin density, $d_{X\sigma}$ the dipole moment of the exchange-hole and its reference electron, approximated according to the Becke-Roussel model,⁷⁸ and $w_i(r)$ represents atomic partitioning weights.

Becke and Johnson⁵⁶ used classical Hirshfeld weightings:⁷⁹

$$w_{i,\text{HC}}(r) = \frac{\rho_i^{\text{at}}(r)}{\sum_n \rho_n^{\text{at}}(r)} \quad (7)$$

where ρ_i^{at} is the sphericalized free atomic density of atom i , weighted by the superposition of all ρ_i^{at} with all atoms n positioned as in the real molecule. The classical Hirshfeld scheme depends on the (arbitrary) choice of the atomic reference densities. Molecules with large ionic character, such as LiF, offer a clear illustration of this dependence. If one uses the typical superposition of neutral atomic densities (i.e., Li⁰ and F⁰), the atomic charges have an absolute value of 0.57. However, a value of 0.98 is obtained when Li⁺ and F[−] densities are considered.⁶³ This arbitrariness can be overcome by using the iterative version of the Hirshfeld partitioning procedure, called Hirshfeld-I.⁶³ In the k th iteration, the weight for atom i is given by

$$w_{i,\text{HI}}^k(r) = \frac{\rho_i^{k-1}(r)}{\sum_n \rho_n^{k-1}(r)} \quad (8)$$

Conveniently, the first iteration can use neutral atomic densities, leading to the classical Hirshfeld charges. Of course, the electronic populations, $N_i = \int w_i(r) \rho(r) dr$, are usually fractional numbers, and the corresponding densities are thus computed according to⁸⁰

$$\rho_i^k = \rho_i^{N_i} = \rho_i^{n+x} = x \cdot \rho_i^{n+1} + (1-x) \cdot \rho_i^n \quad (9)$$

where n is the integer part of N_i and $x = N_i - n$. The partitioning is converged if the electronic populations do not change significantly between two iterations (the convergence criterion was set to a root-mean-square deviation of 0.0005 au). Compared to the rest of the correction, the iterative scheme is computationally demanding, as integration over the entire grid is necessary for each iteration.⁸¹ For this reason, we also report values based on the classical Hirshfeld partitioning.

Finally, the determination of the dispersion coefficients from eqs 3–5 also depends on atomic polarizabilities. We herein follow Becke and Johnson's proposal to exploit the proportionality⁸² between polarizability and volume to estimate the effective atom in molecule (AIM) polarizabilities from tabulated free atomic polarizabilities:⁸³

$$\alpha_i = \frac{\langle r^3 \rangle_i}{\langle r^3 \rangle_{i,\text{free}}} \alpha_{i,\text{free}} = \frac{\int r^3 w_i(r) \rho(r) d^3r}{\int r^3 \rho_{i,\text{free}}(r) d^3r} \alpha_{i,\text{free}} = \frac{V_{i,\text{AIM}}}{V_{i,\text{free}}} \alpha_{i,\text{free}} \quad (10)$$

ii. The Damping. A key component of our dDXDM correction is the damping factor b . We showed previously⁴⁶ that the performance of the TT-damping function is improved by the introduction of a second damping function to prevent corrections at covalent distances. In the full TT model,⁷⁷ the attractive potential should give relatively strong contribution at short distances in order to soften the repulsive Born–Mayer potential. In contrast, a correction to DFT necessitates additional damping as density functionals better describe the region of strong density overlap (short-range). We herein introduce a variable, damped b , in which the second damping is intrinsically absorbed as an alternative to our previous model using a Fermi damping function.⁴⁶ In Tang and Toennies' seminal work,⁷⁷ the damping parameter b is also the range parameter of the repulsive Born–Mayer potential and thus depends on the two interacting atoms. Later, the same authors converted b from a constant into a function:⁸⁴ for an arbitrary repulsive potential $V(r)$,

$$b(r) = -\frac{d \ln V(r)}{dr} \quad (11a)$$

Here, we replace the distance dependence by the following form:

$$b(x) = F(x) \cdot b_{ij,\text{asym}} \quad (11)$$

x and $F(x)$ are respectively the damping argument and the function for $b_{ij,\text{asym}}$, the TT-damping factor associated with two separated atoms. $b_{ij,\text{asym}}$ is computed according to the combination rule:^{85,86}

$$b_{ij,\text{asym}} = 2 \frac{b_{ii,\text{asym}} \cdot b_{jj,\text{asym}}}{b_{ii,\text{asym}} + b_{jj,\text{asym}}} \quad (12)$$

The $b_{ii,\text{asym}}$ values are estimated^{87,88} by the square root of the atomic ionization energy $\sqrt{I_i}$ taken from the literature.⁸⁹ Inspired by the approach of Tkatchenko and co-workers,^{45,90} the atom in molecule character is taken into account through a cubic root scaling of the ratio between the free atom and the AIM volume. After introduction of the parameter b_0 , which determines the strength of the correction in the medium range, we arrive at

$$b_{ii,\text{asym}} = b_0 \cdot \sqrt{2I_i} \cdot \sqrt[3]{\frac{V_{i,\text{free}}}{V_{i,\text{AIM}}}} \quad (13)$$

Equation 14 proved to be the most robust form for the damping function⁹¹

$$F(x) = 1 - \frac{2 \arctan(a_0 \cdot x)}{\pi} \quad (14)$$

where the fitted parameter a_0 adjusts the short-range behavior of the correction.

The last element of the correction is the damping argument x

$$x = \text{abs} \left(q_{ij} + q_{ji} - \frac{(Z_i - N_i) \cdot (Z_j - N_j)}{r_{ij}} \right) \frac{N_i + N_j}{N_i * N_j} \quad (15)$$

where Z_i and N_i are the nuclear charge and Hirshfeld population of atom i (*vide supra*), respectively. The overlap population⁹² $q_{ij} = \int w_i(r) w_j(r) \rho(r) dr$ is a covalent bond index, and the fraction term in the parentheses is an ionic bond index.⁹³ The multiplicative factor, $(N_i + N_j)/(N_i \cdot N_j)$, serves to attenuate the damping of $b_{ij,\text{asym}}$ for heavier atoms (containing more electrons). Note that the damping function has an adequate form (i.e., $F(0) = 1$ and $F(\infty) = 0$), given that x is large for atoms near each other and vanishes with increasing r_{ij} distance.

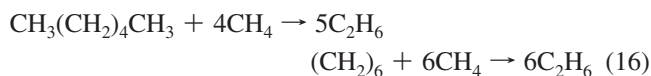
This is the first example for which the damping of an atom pair-wise dispersion correction depends on Hirshfeld (overlap) populations rather than on "critical" or "van der Waals" radii. Our approach is, however, similar in spirit to Slipchenko and Gordon's⁹⁴ overlap-matrix-based formula employed within the framework of the effective fragment potential method.

To summarize, the presented dDXDM correction uses electronic structure information to determine dispersion coefficients and two fitted damping parameters that are the strength of the TT-damping (b_0) and the steepness factor (a_0).

Determination of the Adjustable Parameters

In line with our former work,^{40,46} the chosen fitting procedure ensures a successful treatment of both weak intra- (short-range) and inter- (long-range) molecular interactions. From

a theoretical perspective, typical weakly bound systems, such as rare gas dimers, seem the appropriate choice as a training set. However, the description of rare gas dimers by standard density functionals is not consistent; for instance, PBE overbinds the helium dimer and underbinds the argon dimer (see the Supporting Information). Such behavior is not easily improved by a dispersion correction and highlights that inclusion of rare gas dimers into the training set does not necessarily guarantee a generally improved treatment of weak intra- and intermolecular interactions.^{95,96} In contrast, we and others demonstrated that the large DFT errors in the description of alkane intramolecular interactions (e.g., isomerization energies) are systematic^{9,12} and conveniently reduced by a dispersion correction.^{40,76,97–99} Our recent work, introducing a flexible TT-based correction,⁴⁶ demonstrated that using alkane reaction energies as a training set results in a highly transferable correction, which outperforms others, even for systems well outside the range of the training set (e.g., intermolecular complexes).⁴⁶ Akin to our former fitting procedure, the two parameters (a_0 and b_0) are fitted for each functional as to minimize the mean absolute deviation (MAD) over five reaction energies that are the Pople's isodesmic bond energy separation reaction of *n*-hexane and cyclohexane,



the folding energy of $\text{C}_{22}\text{H}_{46}$, and the isomerization energy of *n*-octane and *n*-undecane to 2,2,3,3-tetramethylbutane and 2,2,3,3,4,4-hexamethylpentane, respectively.

The best fit parameters are given in the Supporting Information for dDXDM (i.e., iterative Hirshfeld weights and terms up to C_{10}), dDXDMc (using classical Hirshfeld weights), dDXDM6 (iterative Hirshfeld weights, only up to C_6), and dDXDM6c (classical Hirshfeld weights and only up to C_6). Short form parenthetical notations that are used in the text refer to the two levels of correction with or without the parentheses (e.g., dDXDM6(c) refers to dDXDM6 and dDXDM6c).

Table S1 and Figure S1 (Supporting Information) illustrate that, for the models including terms up to C_{10} , best fit a_0 and b_0 correlate well with each other. There is also a good correlation between each of the fitted parameters and the repulsive character of the functional, as represented by the error in the methane dimer interaction energy shown in Figure S2 (Supporting Information).^{100,101} In contrast, the C_6 -based corrections show poor (dDXDM6) or even no (dDXDM6c) correlation between a_0 and b_0 . The missing higher order dispersion terms in dDXDM6c are compensated by relatively higher b_0 values.¹⁰² The a_0 parameters adjust accordingly following the repulsive character of the functional to prevent a too strong correction in the short range. These results emphasize the physical relevance of including higher dispersion terms to achieve a more consistent correction.

Test Sets

The robustness of the dDXDM correction is tested on seven illustrative sets featuring both intra- and intermolecular weak interactions, as described hereafter.

Three of the sets assess Pople's isodesmic bond separation equation reactions^{7,8} of saturated hydrocarbons (H, R, and C for chains, rings, and cages, respectively). As in ref 46, B3LYP/6-311+G** geometries and thermal corrections are included, and reference values are derived from experimental heats of formation.¹⁰³

The "intramolecular dispersion interactions in hydrocarbons" (IDHC)⁷⁶ set contains two isomerization reactions (*n*-octane and *n*-undecane to the fully branched isomer), two folding reactions of large hydrocarbon chains ($\text{C}_{14}\text{H}_{30}$ and $\text{C}_{22}\text{H}_{46}$), the dimerization of anthracene, and the hydrogenation reaction of [2.2]paracyclophane to *p*-xylene. Geometries and reference values are taken from ref 76.

The S22⁹⁶ set validates the performance of the correction on noncovalent complexes, while the P76 set test probes peptide conformational energies.¹⁰⁴ P76 contains 76 conformations of five small peptides having aromatic side chains (FGG, GFA, GGF, WG, and WGG). For these two sets, geometries and reference values (estimated CCSD(T)/CBS) are taken from the literature.^{105,106}

The last test set (EX3) exclusively features weak interactions involving heavy atoms in the dimers of pnictogen trihalides (NF_3 , NCl_3 , PCl_3 , PBr_3 , and AsBr_3).¹⁰⁷ Geometries (counterpoise corrected df-MP2/aug-cc-pVTZ) were taken from ref 107. Reference values (estimated CCSD(T)/CBS) were computed at the counterpoise corrected level¹⁰⁸ according to

$$\begin{aligned} E(\text{CCSD(T)/CBS}) &= \text{HF/AVQZ} + \\ &\quad \text{CCSD-F12b/CBS(AVTZ/AVQZ)} + \\ &\quad \text{(T)/CBS(AVDZ/AVTZ)} \quad (17) \end{aligned}$$

where aug-cc-pVDZ, aug-cc-pVTZ, and aug-cc-pVQZ are abbreviated by AVDZ, AVTZ, and AVQZ, respectively. These computations were performed with Molpro2009.1¹⁰⁹ at the F12 level,¹¹⁰ with the HF energy containing the CABS single correction and the triples being based on F12 amplitudes. The *g* functions are omitted in all aug-cc-pVQZ computations, except for the heaviest dimer (i.e., $(\text{AsBr}_3)_2$). The extrapolation functional proposed by Helgaker and co-workers^{111,112} ($E_n^{\text{corr}} = E_{\text{CBS}}^{\text{corr}} + AX^{-3}$ with $X = 2, 3$, and 4 for AVDZ, AVTZ, and AVQZ, respectively) is applied *a posteriori* to the CCSD-F12b and (T) correlation energies.¹¹³ The T1 diagnostic was below 0.02 and the D1 diagnostic¹¹⁴ around 0.04, except for NCl_3 , where $\text{D1} \approx 0.065$ (monomer and dimer) is indicative of a multireference character. The NBr_3 dimer was discarded from the test set due to its $\text{D1} \approx 0.085$ and an unreliable basis-set convergence.

The performance of the dDXDM correction was further examined on four potential energy profiles: (a) the stacked benzene dimer (geometry and reference values taken from refs 115 and 116, respectively), (b) a propane dimer conformation (geometry based on the experimental geometry¹¹⁷ and arranged like in ref 118), (c) a benzene– H_2S complex (geometry and reference from ref 116), and (d) a benzene– H_2O complex (orientation analogous to the benzene– H_2S conformation, with the same benzene geometry¹¹⁹ and the experimental water geometry).¹¹⁷ For b and c, reference values were computed at the counterpoise corrected level.¹⁰⁸

$$E(\text{CCSD(T)/CBS}) = \text{df-MP2/CBS(AVDZ,AVTZ)} + \Delta\text{CCSD(T}^*\text{)-F12b/AVDZ}$$

where $\Delta\text{CCSD(T}^*\text{)-F12b/AVDZ}$ is the difference between df-MP2-F12 and CCSD(T^{*})-F12b evaluated with the aug-cc-pVDZ basis set, and (T^{*}) stands for the perturbative triple corrections improved by scaling by the ratio of df-MP2-F12/df-MP2.¹²⁰

Finally, to ensure that the corrections do not affect covalent bonds, the performance on six representative atomization energies (AE6) and barrier heights (BH6)¹²¹ was investigated. Geometries and reference values were obtained from the Minnesota database collection.¹²² Errors for these two test sets are given in the Supporting Information.

Computational Methods

B97-D and B2PLYP-D computations with the cc-pVTZ basis set^{123–125} were performed with Turbomole 5.10^{126,127} using the resolution of identity (RI-MP2)¹²⁸ with matching auxiliary basis functions¹²⁹ to speed up B2PLYP. M06-2X energies were computed with NWChem 5.1^{130,131} using the “xfine” grid. All of the other computations were performed with a developmental version of Q-Chem 3.2.¹³² The cc-pVTZ basis set^{123–125} was used except for the potential energy curves, for which the larger aug-cc-pVTZ basis set was employed. The energy differences between cc-pVTZ and the larger aug-cc-pVTZ basis set were found to be negligible compared to the error of the method against the reference value¹³³ (e.g., the averaged total MAD for PBE/cc-pVTZ, 4.27 kcal mol⁻¹, differs by only 2%, 0.08 kcal mol⁻¹, from PBE/aug-cc-pVTZ, 4.20 kcal mol⁻¹; see Table S2 Supporting Information).

To ensure a consistent treatment between intra- and intermolecular interaction, no basis set superposition correction was applied (e.g., P76 contains peptide conformations with intramolecular interactions resembling closely those of intermolecular complexes in the S22 test set).

XDM-based corrections were done post-SCF. The iterative Hirshfeld partitioning was implemented using sphericalized restricted–open atomic densities computed on the fly (i.e., functional specific) with a 99/590 Euler–Maclaurin–Lebedev^{134,135} grid. The energy profiles were computed with a 99/302 Euler–Maclaurin–Lebedev grid. Otherwise, the SG1 grid¹³⁶ was used.

Results and Discussion

Figure 1 summarizes the mean absolute deviation for established methods tested on the seven sets described above. The difference between “standard” and “recent” functionals (M06-2X, B97-D, and B2PLYP-D) is significant for all of the test sets (averaged total MAD 5.0 vs 1.5 kcal mol⁻¹). As noted previously,⁴⁶ the performance of the recent functionals on hydrocarbon reaction energies (H, R, C, and IDHC) is significantly better than that of the standard ones (MAD of 3.8 and 12.9 kcal mol⁻¹, respectively), although chemical accuracy has yet to be obtained.

The MADs for the best performing variant of the correction (-dDXDM i.e., iterative Hirshfeld weights and terms up

to C₁₀) are shown in Figure 2a. Note that (un)corrected B2LYP (0.47 B88 + 0.53 HF + 0.73 LYP, same functional contributions as in B2PLYP⁷⁵) is not intended for “real world” applications but provides insight into the good performance of B2PLYP-D. Overall, dDXDM largely improves the parent functionals, yielding low errors. Over the seven corrected functionals tested, the averaged total MAD (TMAD) is 0.9 kcal mol⁻¹ (min 0.74 (PBE0-dDXDM); max 1.11 (BLYP-dDXDM)), significantly lower than for the recent M06-2X, B97-D, and B2PLYP-D (1.5 kcal mol⁻¹, min 1.06 (B2PLYP-D)). The correction improves the IDHC energies for both PBE and HF (MAD of 12.3 and 22.2 kcal mol⁻¹, respectively) to a respectable mean absolute deviation of 1.6 kcal mol⁻¹. B2LYP- and BHHLYP-dDXDM give remarkably low MADs of 0.6 and 0.9 kcal mol⁻¹ (B2PLYP-D gives 1.6 kcal mol⁻¹), while BLYP-dDXDM performs less convincingly (MAD of 3.6 kcal mol⁻¹) for this set. The robustness and range of applicability of dDXDM combined with various functionals is further illustrated by the consistent improvement of alkane BSE reaction energies and weak intermolecular interactions: averaged MADs for the HRC, P76 (relative conformational energies of small peptides), and S22 (intermolecular weak interactions) sets are 1.4, 0.7, and 0.9 kcal mol⁻¹, respectively, corresponding to roughly 10, 50, and 30% of the deviations of the uncorrected values (12.9, 1.3, and 3.2 kcal mol⁻¹). The 0.5 kcal mol⁻¹ averaged MAD for the pure inorganic test set (EX3; vs an uncorrected 3.9 kcal mol⁻¹) is also rewarding. It is worthwhile noting that the proposed corrections do not affect significantly properties such as atomization energies and barrier heights (see the Supporting Information).

PBE0-dDXDM is the most accurate combination presented herein (TMAD of 0.74 kcal mol⁻¹) but dDXDM with the popular B3LYP functional is, as well, very satisfactory (TMAD of 0.82 kcal mol⁻¹). The best corrected GGA, PBE-dDXDM, performs nearly as well as PBE0-dDXDM with a TMAD of 0.84 kcal mol⁻¹. Such a performance is of interest for applications to large systems (or even bulk materials), where hybrid functionals are computationally considerably more demanding. Nevertheless, hybrid functionals, which generally outperform the GGA in many thermochemistry applications, provide the best dDXDM corrected results.

Classical Hirshfeld Partitioning and C₆-Only Dispersion Corrections. The reliability of simpler variants of the correction, i.e., including only terms up to C₆ or using Hirshfeld classical instead of iterative weights, has also been evaluated. The use of the classical Hirshfeld weights is of practical interest, as it is significantly less computationally demanding than the iterative version. In the BJ formalism, C₈/R⁻⁸ and C₁₀/R⁻¹⁰ terms are relatively inexpensive but have non-negligible contributions to the interaction energy at short internuclear separations.^{49,58,102} A comparison with the C₆ truncation is thus of theoretical relevance.

Figure 2a (dDXDM) and b (dDXDM6) reveal that the BSE of alkane cages, the IDHC, and the EX3 test sets are most affected by the truncation. Whereas the first two sets are characterized by a high number of short-range interactions, the effect in the EX3 interaction energies is more difficult

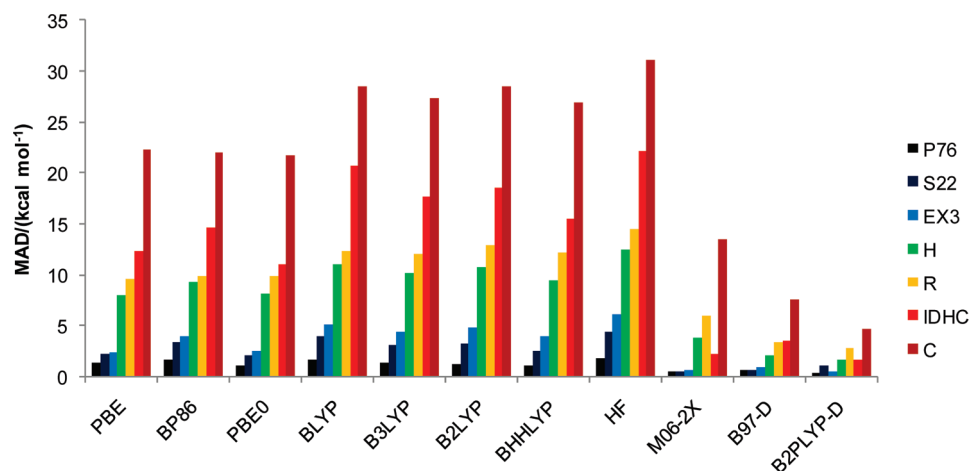


Figure 1. Performance for commonly used functionals: Mean absolute deviations for binding energies for noncovalent complexes (S22 and EX3); relative conformational energies of five small peptides (P76); and bond separation energies over hydrocarbon chains (H), rings (R), and cages (C) and for reaction energies of the test set “intramolecular dispersion interactions” (IDHC) using the cc-pVTZ basis set.

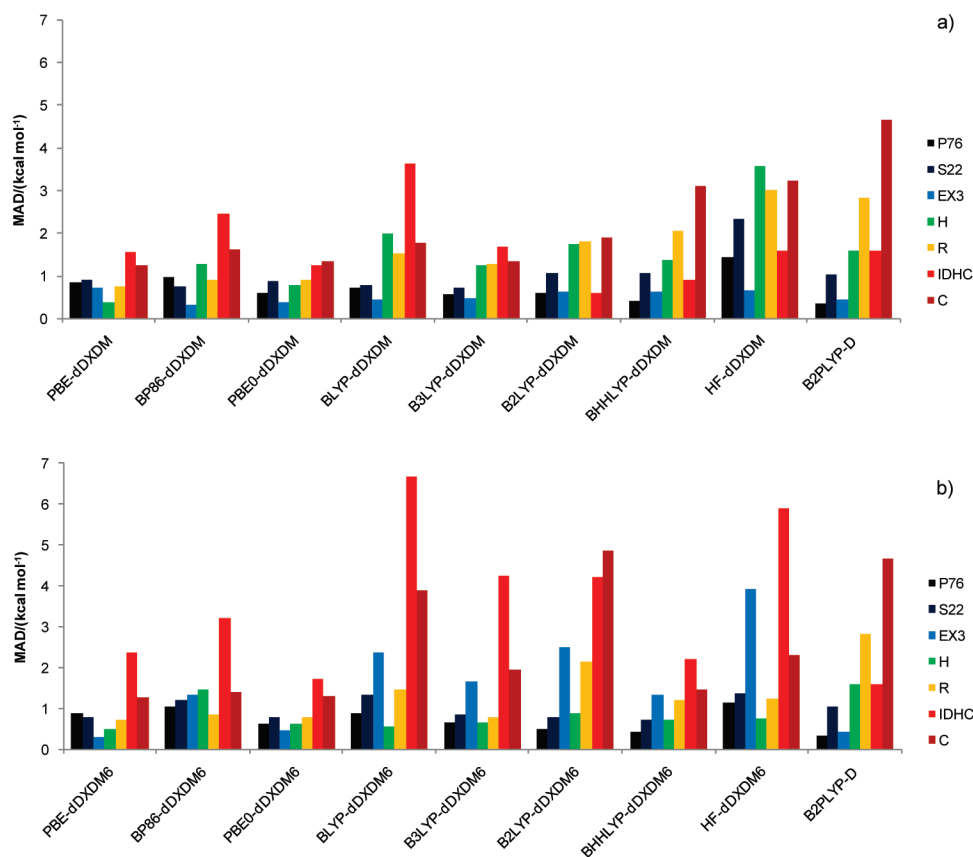


Figure 2. Performance for the iterative Hirshfeld-distributed dispersion coefficients up to C_{10} (a) and up to C_6 (b): Mean absolute deviations for binding energies for noncovalent complexes (S22 and EX3); relative conformational energies of five small peptides (P76); and bond separation energies over hydrocarbon chains (H), rings (R), and cages (C) and for reaction energies of the test set “intramolecular dispersion interactions” (IDHC) using the cc-pVTZ basis set. B2PLYP-D serves as an “internal standard”.

to interpret. Overall, only the combinations of dDXDM6 with PBE, PBE0, and BHLYP match the dDXDM results closely.

For the higher-order multipole expansion, classical Hirshfeld weights result in larger errors than the iterative procedure (Figure 3). With an increase in averaged MAD from 0.9 (dDXDM) to 1.5 kcal mol⁻¹ (dDXDMc), the S22 test set is the most representative of the classical partitioning

limitation (underestimation of ionic characters).¹³⁷ As an example, the C_6 -(PBE) $O\cdots O/H\cdots H$ dispersion coefficients for the water dimer are 12.6/2.5 with classical Hirshfeld weights, compared to 21.2/0.9 with the iterative procedure. The key difference arises from the ionic bond index appearing in eq 13. The index for the $O\cdots O$ atom pair is 0.014 while using atomic densities (classical partitioning) and 0.15 after the iterative scheme. This difference translates

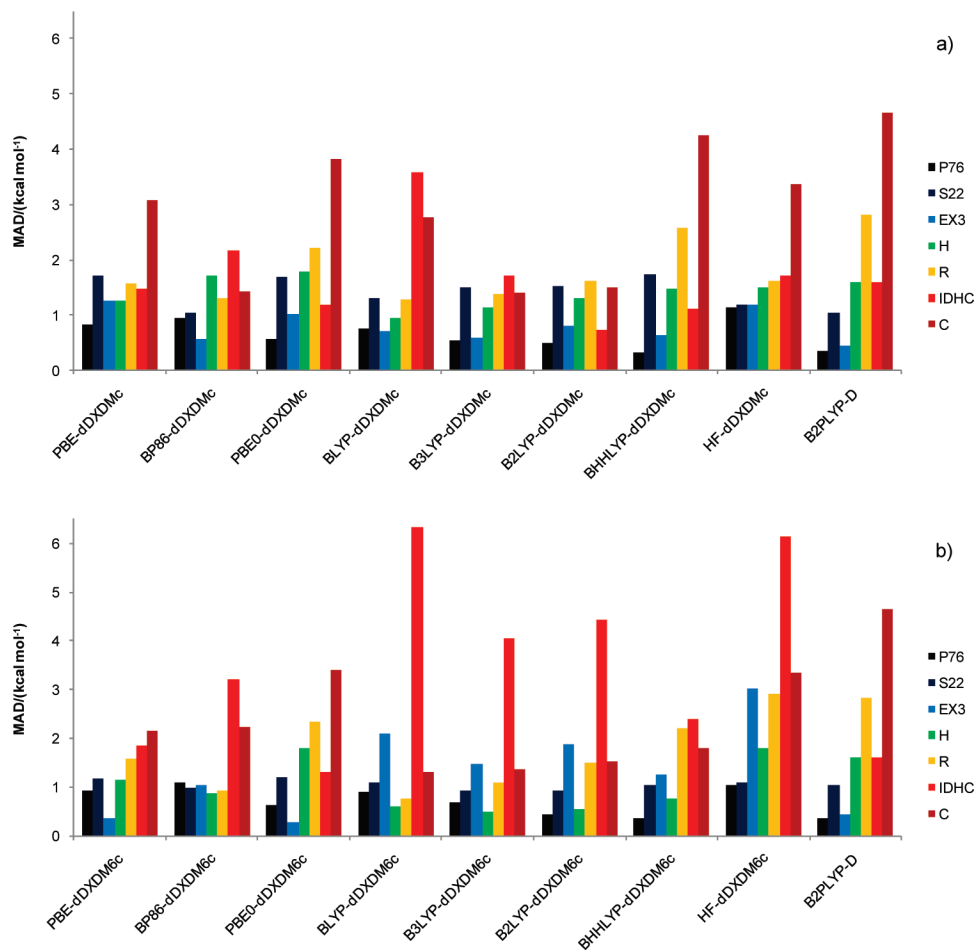


Figure 3. Performance for the classical Hirshfeld distributed dispersion coefficients up to C_{10} (a) and up to C_6 (b): Mean absolute deviations for binding energies for noncovalent complexes (S22 and EX3); relative conformational energies of five small peptides (P76); and bond separation energies over hydrocarbon chains (H), rings (R), and cages (C) and for reaction energies of the test set “intramolecular dispersion interactions” (IDHC) using the cc-pVTZ basis. B2PLYP-D serves as an “internal standard”.

into a strong/weak damping when iterative/classical Hirshfeld charges are used. As DFT methods correctly account for interaction energy between strongly polarized fragments (e.g., H bonds), higher iterative Hirshfeld charges (i.e., strong damping, small corrections) are better suited. In contrast, HF that systematically underestimates electrostatic interactions benefits from the larger dispersion corrections associated with the use of classical Hirshfeld weights. It is thus not surprising that Hartree–Fock gives its best results when combined with dDXDMc (TMAD of $1.3 \text{ kcal mol}^{-1}$, $\text{MAD}(\text{S22}) = 1.18 \text{ kcal mol}^{-1}$) and that HF-dDXDM is the least accurate variant (TMAD of $2.01 \text{ kcal mol}^{-1}$, $\text{MAD}(\text{S22}) = 2.32 \text{ kcal mol}^{-1}$). HF-dDXDMc could thus be a general alternative to the recent refined HF-D approach, which has been proven to be successful for intermolecular interactions.⁴⁸

For the reasons given above, the classical Hirshfeld partitioning performs better on the S22 set when terms only up to C_6 are included (see Figure 3b): excluding higher dispersion corrections attenuates the overcorrections of polar interactions. With TMADs below $1.0 \text{ kcal mol}^{-1}$, B3LYP-dDXDM6c and BHLYP-dDXDM6c represent attractive alternatives to avoid the iterative scheme. As for the GGAs, PBE-dDXDM6c and BP86-dDXDMc are the most consistent over the seven sets tested (TMAD of 1.12 and $1.14 \text{ kcal mol}^{-1}$, respectively). Comparisons of B2LYP-dDXDM6(c)

to B2PLYP-D and B2LYP-dDXDM demonstrate that the C_6/R^6 -dispersion terms are not sufficient to correct B2LYP errors in the EX3 and IDHC sets. Including either higher dispersion terms semiempirically as in B2LYP-dDXDM(c) or adding a fraction of PT2 energy to give B2PLYP-D is crucial for these two test sets. Apart from those, B2LYP-dDXDM6(c) performs similarly to B2PLYP-D, even improving alkane BSE energies. Corrected B2LYP and B3LYP also tend to perform the same. The similarity relies on the fitting procedure used to determine the empirical parameters of both, B3LYP and B2PLYP.

Interaction Energy Profiles. Figure 4 shows potential energy curves of complexes typically underbound at the (hybrid)-GGA levels (stacked benzene dimer (a), propane dimer (b), and the benzene complex with water (c) and hydrogen sulfide (d)). The hybrid-meta-GGA M06-2X offers substantial improvement for the benzene– H_2S complex but under- and overbinds the stacked benzene dimer conformation and the water–benzene complex, respectively. PBE-dDXDM, B3LYP-dDXDM, and, to a lesser extent, B2PLYP-D overbind all four complexes, while the dDXDM6c corrections provide significantly better results for these weakly bound complexes (*vide infra*). Since B2PLYP-D suffers greatly from basis set superposition and incompleteness

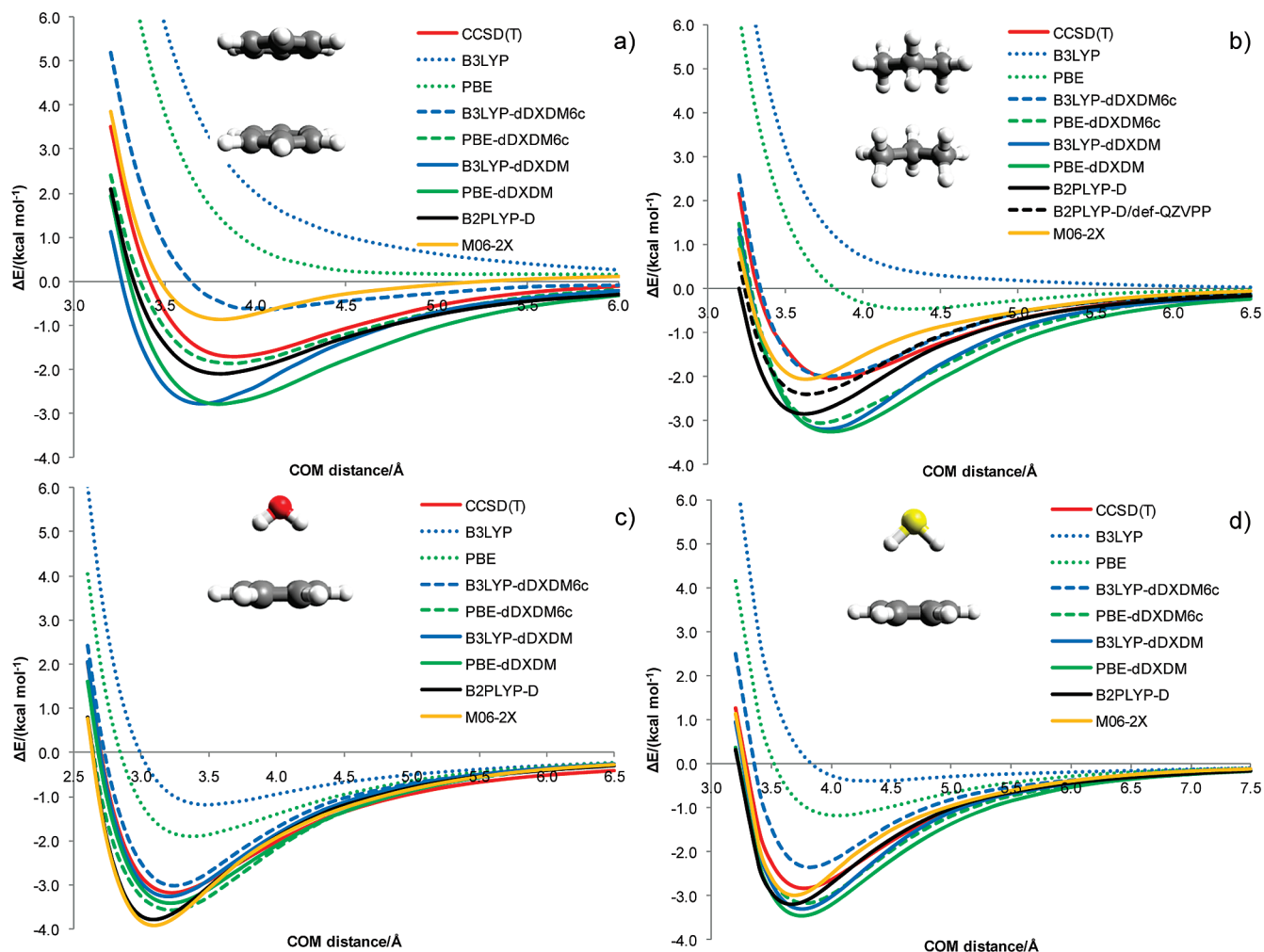


Figure 4. Interaction energy (ΔE) profiles for the (a) stacked benzene dimer, (b) propane dimer, (c) benzene–H₂O complex, and (d) benzene–H₂S complex. CCSD(T) references for a and d are taken from ref 116, while b and d are computed (see Test Sets). If not stated otherwise, density functional computations were performed with the aug-cc-pVTZ basis set.

Table 1. MAD (in kcal mol⁻¹) and Mean Absolute Relative Deviation (in percent) over All 67 Points of Figure 4

	MAD	mean absolute relative deviation
B3LYP	2.67	357.2
PBE	1.69	222.0
B3LYP-dDXDM6c	0.45	56.8
PBE-dDXDM6c	0.39	58.6
B3LYP-dDXDM	0.49	54.9
PBE-dDXDM	0.59	69.9
B2PLYP-D	0.47	81.8
M06-2X	0.41	75.2

errors,^{75,76} both B2PLYP-D/aug-cc-pVTZ and B2PLYP-D/def-QZVPP energy curves are reported for the propane dimer. As expected, the accuracy of the energy curve is drastically improved with the large def-QZVPP basis set.

The MAD and mean absolute relative deviation over all 67 points associated with the four potential energy curves are given in Table 1. Figure 5, on the other hand, displays the error in the propane dimer interaction energy. With the exception of PBE-dDXDM, all dispersion-corrected methods have MADs between 0.4 and 0.5 kcal mol⁻¹. PBE-dDXDM6c is the most accurate combination (MAD 0.39 kcal mol⁻¹). The distinctive performance of the current corrections is further emphasized by the remarkably low error in

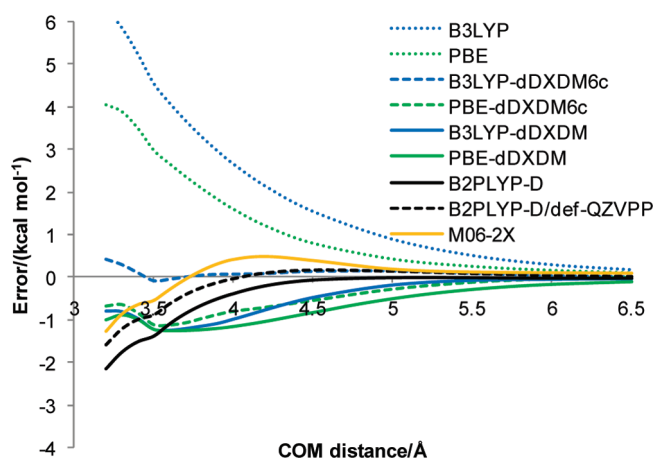


Figure 5. Errors (with respect to estimated CCSD(T)/CBS) in DFT interaction energies for the propane dimer.

both the short (i.e., repulsive wall) and long range of Figure 5. Overall, the error range spans between 55% (B3LYP-dDXDM) and 70% (PBE-dDXDM), thereby outperforming M06-2X (75%) and B2PLYP-D (80%) (Table 1).

As discussed earlier, DFT binding energies of the rare-gas dimers cannot be easily corrected by a dispersion correction. Nevertheless, those archetypical systems represent

a challenging set for testing the robustness of our correction, and their interaction energy profiles (i.e., helium, neon, and argon homodimers) are, for this reason, given in the Supporting Information (Figure S3). It can be seen that, whereas PBE overbinds after correction and is thus less satisfactory for rare-gas dimers, our corrected B3LYP and HF interaction energies compare well with M06-2X or B2PLYP-D, two other generally well performing approximations.

Conclusions

We have presented an improved scheme for computing system-dependent dispersion coefficients and damping parameters for a correction to density functional theory. The dispersion coefficients are evaluated exploiting the XDM formalism of Becke and Johnson^{37,55–59} and are distributed among the atoms according to a(n) (iterative)⁶³ Hirshfeld⁷⁹ partitioning. The universal damping function of Tang and Toennies⁷⁷ is used with a damping factor depending on Hirshfeld (overlap) populations and charges as well as on two adjustable parameters. In addition to the fitted parameters and the density-based information, only free atomic polarizabilities and ionization energies are needed. Hence, the dDXDM correction is applicable to all elements of the periodic table and is easily combined with every density functional. This flexibility permits choosing a functional on the basis of its performance for properties not dominated by weak interactions (e.g., spin states and barrier heights), while still correcting any failures for weak interactions. The analysis of 30 (dispersion corrected) density functionals on 145 systems reveals that dDXDM(6c) largely reduces the error of the parent functionals for both inter- and intramolecular interactions. PBE0-dDXDM and PBE-dDXDM are the best performing hybrid-GGA and GGA, respectively, outperforming M06-2X and B2PLYP-D. The use of B3LYP-dDXDM is recommended as well, and it gives the second best overall performance.

Acknowledgment. C.C. acknowledges the Sandoz Family Foundation, Swiss NSF Grant 200021_121577/1, and EPFL for financial support. We are grateful to Q-Chem Inc. for providing the source code and for helpful discussions with Drs. Zhengting Gan and Jing Kong. The authors also thank Drs. Gabor Csonka and Alexandre Tkatchenko for stimulating conversations.

Supporting Information Available: Optimal a_0 and b_0 values for all corrections are given. (Corrected) PBE/aug-cc-pVTZ results are listed. Absolute and reaction energies for all systems and the detailed values used for estimating the CCSD(T)/CBS limit of the propane dimer and the benzene–H₂O complex are provided. Cartesian coordinates for the H, R, C test set, the propane dimer, and the benzene–H₂O complex as well as details on the rare gas dimers, barrier heights, and atomization energy test sets are also available.

This material is provided free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- (2) Meijer, E. J.; Sprik, M. *J. Chem. Phys.* **1996**, *105*, 8684–8689.
- (3) Wu, X.; Vargas, M. C.; Nayak, S.; Lotrich, V.; Scoles, G. *J. Chem. Phys.* **2001**, *115*, 8748–8757.
- (4) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149–5155.
- (5) Dabkowska, I.; Gonzalez, H. V.; Jurecka, P.; Hobza, P. *J. Phys. Chem. A* **2005**, *109*, 1131–1136.
- (6) Bashford, D.; Chothia, C.; Lesk, A. M. *J. Mol. Biol.* **1987**, *196*, 199–216.
- (7) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. *J. Am. Chem. Soc.* **1970**, *92*, 4796–4801.
- (8) Pople, J. A.; Radom, L.; Hehre, W. J. *J. Am. Chem. Soc.* **1971**, *93*, 289–300.
- (9) Wodrich, M. D.; Corminboeuf, C.; Schleyer, P. v. R. *Org. Lett.* **2006**, *8*, 3631–3634.
- (10) Wodrich, M. D.; Corminboeuf, C.; Schreiner, P. R.; Fokin, A. A.; Schleyer, P. v. R. *Org. Lett.* **2007**, *9*, 1851–1854.
- (11) Schreiner, P. R. *Angew. Chem., Int. Ed.* **2007**, *46*, 4217–4219.
- (12) Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460–4464.
- (13) Lotrich, V. F.; Szalewicz, K. *J. Chem. Phys.* **1997**, *106*, 9668–9687.
- (14) Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2005**, *123*, 214103.
- (15) Podeszwa, R.; Szalewicz, K. *J. Chem. Phys.* **2007**, *126*, 194101.
- (16) Aeberhard, P. C.; Arey, J. S.; Lin, I. C.; Rothlisberger, U. *J. Chem. Theory Comput.* **2008**, *5*, 23–28.
- (17) Cascella, M.; Lin, I.-C.; Tavernelli, I.; Rothlisberger, U. *J. Chem. Theory Comput.* **2009**, *5*, 2930–2934.
- (18) Lin, I.-C.; Coutinho-Neto, M. D.; Felsenheimer, C.; Lilienfeld, O. A. v.; Tavernelli, I.; Rothlisberger, U. *Phys. Rev. B* **2007**, *75*, 205131.
- (19) Lilienfeld, O. A. v.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.
- (20) Lilienfeld, O. A. v.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. B* **2005**, *71*, 195119.
- (21) Mackie, I. D.; DiLabio, G. A. *J. Phys. Chem. A* **2008**, *112*, 10968–10976.
- (22) Nilsson Lill, S. O. *J. Phys. Chem. A* **2009**, *113*, 10321–10326.
- (23) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715–2719.
- (24) Zhao, Y.; Truhlar, D. *Theor. Comput. Model.* **2008**, *120*, 215–241.
- (25) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908–6918.
- (26) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 4209–4212.
- (27) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.

- (28) Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673–2677.
- (29) Conway, A.; Murrell, J. N. *Mol. Phys.* **1974**, *27*, 873–878.
- (30) Wagner, A. F.; Das, G.; Wahl, A. C. *J. Chem. Phys.* **1974**, *60*, 1885–1891.
- (31) Hepburn, J.; Scoles, G.; Penco, R. *Chem. Phys. Lett.* **1975**, *36*, 451–456.
- (32) Ahlrichs, R.; Penco, R.; Scoles, G. *Chem. Phys.* **1977**, *19*, 119–130.
- (33) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.
- (34) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (35) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.
- (36) Zimmerli, U.; Parrinello, M.; Koumoutsakos, P. *J. Chem. Phys.* **2004**, *120*, 2693–2699.
- (37) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.
- (38) Ducere, J.-M.; Cavallo, L. *J. Phys. Chem. B* **2007**, *111*, 13124–13134.
- (39) Olasz, A.; Vanommeslaeghe, K.; Krishtal, A.; Veszpremi, T.; Alsenoy, C. V.; Geerlings, P. *J. Chem. Phys.* **2007**, *127*, 224105.
- (40) Wodrich, M. D.; Jana, D. F.; Schleyer, P. v. R.; Corminboeuf, C. *J. Phys. Chem. A* **2008**, *112*, 11495–11500.
- (41) Murdachaew, G.; de Gironcoli, S.; Scoles, G. *J. Phys. Chem. A* **2008**, *112*, 9993–10005.
- (42) Jurecka, P.; Cerný, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555–569.
- (43) Krishtal, A.; Vanommeslaeghe, K.; Olasz, A.; Veszpremi, T.; Alsenoy, C. V.; Geerlings, P. *J. Chem. Phys.* **2009**, *130*, 174101.
- (44) Liu, Y.; Goddard, W. A. *Mater. Trans.* **2009**, *50*, 1664–1670.
- (45) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.
- (46) Steinmann, S. N.; Csonka, G.; Corminboeuf, C. *J. Chem. Theory Comput.* **2009**, *5*, 2950–2958.
- (47) Pernal, K.; Podeszwa, R.; Patkowski, K.; Szalewicz, K. *Phys. Rev. Lett.* **2009**, *103*, 4.
- (48) Podeszwa, R.; Pernal, K.; Patkowski, K.; Szalewicz, K. *J. Phys. Chem. Lett.* **2009**, *1*, 550–555.
- (49) Sato, T.; Nakai, H. *J. Chem. Phys.* **2009**, *131*, 224104.
- (50) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2009**, *5*, 719–727.
- (51) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- (52) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2010**, *6*, 1081–1088.
- (53) Grimme, S.; Diedrich, C.; Korth, M. *Angew. Chem., Int. Ed.* **2006**, *45*, 625–629.
- (54) Grimme, S.; Steinmetz, M.; Korth, M. *J. Chem. Theory Comput.* **2007**, *3*, 42–45.
- (55) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 154104.
- (56) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.
- (57) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 014104.
- (58) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2006**, *124*, 174104.
- (59) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 154108.
- (60) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 124108.
- (61) Proynov, E.; Gan, Z.; Kong, J. *Chem. Phys. Lett.* **2008**, *455*, 103–109.
- (62) Kong, J.; Gan, Z.; Proynov, E.; Freindorf, M.; Furlani, T. R. *Phys. Rev. A* **2009**, *79*, 042510.
- (63) Bultinck, P.; Alsenoy, C. V.; Ayers, P. W.; Carbo-Dorca, R. *J. Chem. Phys.* **2007**, *126*, 144111.
- (64) Krishtal, A.; Senet, P.; Van Alsenoy, C. *J. Chem. Theory Comput.* **2008**, *4*, 2122–2129.
- (65) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (66) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1986**, *33*, 8800–8802.
- (67) Perdew, J. P.; Yue, W. *Phys. Rev. B* **1989**, *40*, 3399.
- (68) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (69) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
- (70) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (71) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (72) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (73) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (74) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (75) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (76) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.
- (77) Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **1984**, *80*, 3726–3741.
- (78) Becke, A. D.; Roussel, M. R. *Phys. Rev. A* **1989**, *39*, 3761.
- (79) Hirshfeld, F. L. *Theo. Comput. Mod.* **1977**, *44*, 129–138.
- (80) Yang, W.; Zhang, Y.; Ayers, P. W. *Phys. Rev. Lett.* **2000**, *84*, 5172.
- (81) Lowering the convergence threshold and using an improved guess would decrease the number of iterations. The improved guess is expected to be especially efficient for geometry optimization, where partial charges do not vary a lot between two steps.
- (82) Brinck, T.; Murray, J. S.; Politzer, P. *J. Chem. Phys.* **1993**, *98*, 4305–4306.
- (83) Miller, T. M. In *CRC Handbook of Chemistry and Physics*, 90th ed.; Taylor & Francis Group: London.
- (84) Tang, K. T.; Toennies, J. P. *Surf. Sci.* **1992**, *279*, L203–L206.
- (85) Sheng, X. W.; Li, P.; Tang, K. T. *J. Chem. Phys.* **2009**, *130*, 174310.
- (86) Bohm, H.-J.; Ahlrichs, R. *J. Chem. Phys.* **1982**, *77*, 2028–2034.
- (87) Douketis, C.; Scoles, G.; Marchetti, S.; Zen, M.; Thakkar, A. J. *J. Chem. Phys.* **1982**, *76*, 3057–3063.

- (88) Tang, K. T.; Toennies, J. P.; Yiu, C. L. *Phys. Rev. Lett.* **1995**, *74*, 1546.
- (89) Martin, W. C.; Musgrove, A.; Kotochigova, S.; Sansonetti, J. E. In *Physical Reference Data, NIST Standard Reference Database Number 111*; National Institute of Standards and Technology: Gaithersburg, MD, 2003.
- (90) Tkatchenko, A.; Robert, A.; DiStasio, J.; Head-Gordon, M.; Scheffler, M. *J. Chem. Phys.* **2009**, *131*, 094106.
- (91) Different functionals, different order of multipole expansion, classical/iterative Hirshfeld partitioning.
- (92) Mayer, I.; Salvador, P. *Chem. Phys. Lett.* **2004**, *383*, 368–375.
- (93) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1841–1846.
- (94) Slipchenko, L. V.; Gordon, M. S. *Mol. Phys.* **2009**, *107*, 999–1016.
- (95) Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1624–1626.
- (96) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- (97) Karton, A.; Tarnopolsky, A.; Lamère, J.-F. O.; Schatz, G. C.; Martin, J. M. L. *J. Phys. Chem. A* **2008**, *112*, 12868–12886.
- (98) Shamov, G. A.; Budzelaar, P. H. M.; Schreckenbach, G. *J. Chem. Theory Comput.*, *6*, 477–490.
- (99) Song, J.-W.; Tsuneda, T.; Sato, T.; Hirao, K. *Org. Lett.* **2010**, *12*, 1440–1443.
- (100) Johnson, E. R.; Mori-Sanchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 204112.
- (101) A detailed analysis of a correlation of DFT-errors for reaction energies with failures in the short-range potential energy will be reported elsewhere.
- (102) Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **2003**, *118*, 4976–4983.
- (103) Afeefy, H. Y.; Liebman, J. F.; Stein, S. E. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; National Institute of Standards and Technology: Gaithersburg MD.
- (104) Valdes, H.; Pluhackova, K.; Pitonak, M.; Rezac, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2747–2757.
- (105) Rezac, J.; Jurecka, P.; Riley, K. E.; Cerny, J.; Valdes, H.; Pluhackova, K.; Berka, K.; Øezàè, T.; Pitonak, M.; Vondrášek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261–1270.
- (106) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104–5.
- (107) Moilanen, J.; Ganesamoorthy, C.; Balakrishna, M. S.; Tuononen, H. M. *Inorg. Chem.* **2009**, *48*, 6740–6747.
- (108) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.
- (109) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *Molpro2009.1*; Cardiff University: Cardiff, U. K., 2009.
- (110) Adler, T. B.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2007**, *127*, 221106.
- (111) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.
- (112) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.
- (113) Hill, J. G.; Peterson, K. A.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2009**, *131*, 194105.
- (114) Janssen, C. L.; Nielsen, I. M. B. *Chem. Phys. Lett.* **1998**, *290*, 423–430.
- (115) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200–10207.
- (116) Sherrill, C. D.; Takatani, T.; Hohenstein, E. G. *J. Phys. Chem. A* **2009**, *113*, 10146–10159.
- (117) Johnson, R. D., III. In *NIST Computational Chemistry Comparison and Benchmark Database*, 14, Sept 2006 ed. <http://cccbdb.nist.gov/> (accessed Jun 2010).
- (118) Riley, K. E.; Pitonak, M.; Cerny, J.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *6*, 66–80.
- (119) Gauss, J.; Stanton, J. F. *J. Phys. Chem. A* **2000**, *104*, 2865–2868.
- (120) Marchetti, O.; Werner, H.-J. *J. Phys. Chem. A* **2009**, *113*, 11580–11585.
- (121) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 8996–8999.
- (122) Minnesota Database Collection, http://t1.chem.umn.edu/misc/database_group/database_therm_bh/ (accessed Jun 2010).
- (123) Thom, H.; Dunning, J. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (124) Woon, D. E.; Thom, H.; Dunning, J. *J. Chem. Phys.* **1993**, *98*, 1358–1371.
- (125) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Thom, H.; Dunning, J. *J. Chem. Phys.* **1999**, *110*, 7667–7676.
- (126) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.
- (127) Ahlrichs, R. *TURBOMOLE V5.10*; University of Karlsruhe: Karlsruhe, Germany, 2008.
- (128) Weigend, F.; Häser, M. *Theor. Comput. Model.* **1997**, *97*, 331–340.
- (129) Weigend, F.; Kohn, A.; Hattig, C. *J. Chem. Phys.* **2002**, *116*, 3175–3183.
- (130) Kendall, R. A.; Aprà, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Commun.* **2000**, *128*, 260–283.
- (131) Bylaska, E. J.; Govind, W. A. d. J. N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Apra, E.; Windus, T. L.; Hammond, J.; Nichols, P.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P. -D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Wu, Q.; Van Voorhis, T.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, T.; Dachsel, H.; Deegan, M.; Dylla, K.; Elwood, D.; Glendenning, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.;

- Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem*, 5.1 ed.; Pacific Northwest National Laboratory: Richland, WA, 2007.
- (132) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A., Jr.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Voorhis, T. V.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L., III; Zhang, W.; Bell, A. T.; Chakraborty, A. K. *Phys. Chem. Chem. Phys.* **2006**, 8, 3172–3191.
- (133) Grimme, S.; Steinmetz, M.; Korth, M. *J. Org. Chem.* **2007**, 72, 2118–2126.
- (134) Murray, C. W.; Handy, N. C.; Laming, G. J. *Mol. Phys.* **1993**, 78, 997–1014.
- (135) Lebedev, V. I.; Laikov, D. N. *Dokl. Math.* **1999**, 59, 477–481.
- (136) Gill, P. M. W.; Johnson, B. G.; Pople, J. A. *Chem. Phys. Lett.* **1993**, 209, 506–512.
- (137) Davidson, E. R.; Chakravorty, S. *Theor. Comput. Model.* **1992**, 83, 319–330.

CT1001494

Ring Current Model and Anisotropic Magnetic Response of Cyclopropane

Raphaël Carion and Benoît Champagne

Laboratoire de Chimie Théorique, Unité de Chimie Physique Théorique et Structurale, Facultés Universitaires Notre-Dame de la Paix, rue de Bruxelles, 61, B-5000 Namur, Belgium

Guglielmo Monaco* and Riccardo Zanasi

Dipartimento di Chimica dell'Università degli Studi di Salerno, via Ponte don Melillo, 84084 Fisciano (SA), Italy

Stefano Pelloni* and Paolo Lazzeretti

Dipartimento di Chimica dell'Università degli Studi di Modena, via Campi 183, 41100 Modena, Italy

Received April 1, 2010

Abstract: Three-dimensional models of the quantum mechanical current density, induced in the electron cloud of the cyclopropane molecule by a uniform magnetic field applied either along the C_3 or the C_2 symmetry axes (indicated by $B_{||}$ and B_{\perp} , respectively), have been constructed via extended calculations. These models of near Hartree–Fock quality, previously shown to provide a good agreement between computed and observed values of magnetic tensors, have been used to interpret the magnitude of the diagonal components of susceptibility (χ), nuclear shielding of carbon (σ^C) and hydrogen (σ^H), and shielding at the center of mass (σ^{CM}). The source of the exceptionally large in-plane component σ_{\perp}^{CM} , dominating the anomalous average σ_{av}^{CM} , is shown to be a strong delocalized current flowing around the methylene moieties and the noncyclic CH_2 – CH_2 fragment. The total current strength for a magnetic field applied in the direction of a C_2 symmetry axis is 15.7 nA/T, approximately 1.5 times larger than that calculated for $B_{||}$. The largest component of the susceptibility is instead the out-of-plane $\chi_{||}$, which depends on the intensity of the σ -electron currents and on the entire area enclosed within the loops that they form about the C_3 axis, all over its length. In a magnetic field perpendicular to the plane of the carbon atoms, both H and C nuclei sit inside diatropic whirlpools, flowing within the sp^3 hybrid orbital which form the C–H bonds and extending for several bohrs above and below the σ_h plane. The average values and the anisotropy of carbon and proton shieldings are strongly biased by the diamagnetic shift of the out-of-plane tensor components partially determined by these vortices. The current density model of cyclopropane is revised according to these findings.

1. Introduction

The peculiar electronic structure of cyclopropane has prompted many investigations of its magnetic properties.^{1–11} Indeed,

* Corresponding authors. E-mail: gmonaco@unisa.it and Stefano.Pelloni@unimore.it.

starting from the observation that in polycyclic aromatic hydrocarbons the additive rules for the average magnetizability $\chi_{av} = (1/3)(\chi_{xx} + \chi_{yy} + \chi_{zz})$ fail, the nonadditive part of the magnetizability tensor χ^{nonloc} has been considered an indicator of electron delocalization.¹² Actually, as the individual components of χ cannot be measured for mol-

ecules in disordered phase, delocalization was more often discussed in terms of χ_{av} and of the magnetic anisotropy $\Delta\chi = \chi_{cc} - (1/2)(\chi_{aa} + \chi_{bb})$ in the system of (a, b, c) principal axes.¹²

With the advent of NMR spectroscopy, the most used delocalization probe in polycyclics became the chemical shift $\delta^I = \sigma_{av}^{ref} - \sigma_{av}^I$ with respect to a reference compound, where $\sigma_{av}^I = (1/3)(\sigma_{xx}^I + \sigma_{yy}^I + \sigma_{zz}^I)$ is the average magnetic shielding of nucleus I . Thanks to the development of powerful ab initio codes, accurate theoretical values of nuclear magnetic shielding and chemical shift are easily available, and the negative of the chemical shift of a ghost atom placed in a suitably defined “ring center”, referred to as nucleus independent chemical shift (NICS), has also been widely used as a measure of magnetotropy in connection with local, or nonlocal, effects.^{13,14} Arguably, if the largest component β_{cc} of a magnetic tensor $\beta \equiv \chi, \sigma^I$ has a dominant nonlocal component, one expects $3\beta_{av}^{nonloc} \approx \beta_{cc}^{nonloc} \approx \Delta\beta^{nonloc}$. For molecules satisfying this condition, isotropic values and anisotropies should lead to analogous conclusions concerning the amount of electron delocalization on the magnetic criterion.

In 1952 the large χ_{av} of cyclopropane was first interpreted in terms of a ring current model (RCM),¹ widely discussed by other authors.^{3,6–9,11} RCMs have also been examined to rationalize the upfield proton chemical shifts of cyclopropane and its derivatives.^{2,4}

According to the Biot–Savart law (BSL), the π -ring current induced by an external magnetic field B_z at right angles to the plane of benzene reinforces the applied field at the site of the protons, which lie beyond the ring current loop, thus causing a paramagnetic shift of the out-of-plane component σ_{zz}^H of the proton shielding. On the other hand, a ring current would diminish the applied B_z at the protons of cyclopropane, which lie inside the circuit, and thus increases σ_{zz}^H .^{2–4,15}

A “ring current involving cyclic σ -electron delocalization among the three carbon atoms” is explicitly referred to by Dale Poulter et al.,⁷ claiming that it would qualitatively explain the anisotropy of a cyclopropyl group discussed by several workers.^{2,16–18} The model chosen by Dale Poulter et al. “considers the effect of electrons precessing in a circle which circumscribes the ring”, with radius 0.88 Å.⁷ The interpretation of cyclopropane magnetic response in term of ring currents was strongly advocated by Dewar,¹⁵ who proposed the much discussed¹⁹ and still debated concept of σ -aromaticity.²⁰

Benson and Flygare⁵ found it surprising that the experimental value of $\chi_{av} = -39.2 \times 10^{-6}$ erg G⁻² mol⁻¹ in cyclopropane is considerably larger than the corresponding value -28.6×10^{-6} in cyclopropene, whereas the estimated $\Delta\chi = -10.0 \times 10^{-6}$ erg G⁻² mol⁻¹ in cyclopropane is significantly smaller than $\Delta\chi = -17.0 \pm 0.5 \times 10^{-6}$ in cyclopropene.⁵ Therefore, χ_{av} and $\Delta\chi$ would seem to yield opposite orders for the amount of delocalization in these molecules.

Moreover, the constitutive corrections for rings, -3.2 and -4.1 ppm erg G⁻² mol⁻¹ estimated by these authors for the cyclopropane and cyclopropene, respectively,⁵ are $\approx 8.2\%$

and $\approx 14\%$ of the experimental average susceptibility. From the atomic Pascal terms $\chi_C = -6.00$ and $\chi_H = -2.93$, recently reported by Bain and Berry,²¹ one obtains for cyclopropane $\chi_{av}^{nonloc} = -39.2 - 3 \times (-6.00) - 6 \times (-2.93) = -3.6$ ppm erg G⁻² mol⁻¹, quite close to the Benson and Flygare constitutive correction for rings.⁵ From the experimental anisotropy of cyclopropane¹⁰ $\Delta\chi = -11.6$ ppm erg G⁻² mol⁻¹, assuming the contribution $\Delta\chi^{loc} = 1.6 \pm 0.8$ for an sp³ carbon,^{22,23} one obtains $\Delta\chi_{nonloc} = -16.4$ ppm erg G⁻² mol⁻¹. Therefore, in cyclopropane $3\chi_{av}^{nonloc}$ differs from $\Delta\chi_{nonloc}$ by 5.6 cgs ppm units, an amount well above experimental errors, which could be interpreted in terms of electron delocalization enhancing $\chi_{||}$ but also affecting the in-plane component χ_{\perp} .

The nonlocal contributions to the out- and in-plane components of the susceptibility tensor of cyclopropane, in ppm erg G⁻² mol⁻¹, are easily evaluated from these data:

$$\chi_{||}^{nonloc} = \chi_{av}^{nonloc} + \frac{2}{3}\Delta\chi_{nonloc} = -14.5$$

$$\chi_{\perp}^{nonloc} = \chi_{av}^{nonloc} - \frac{1}{3}\Delta\chi_{nonloc} = 1.9$$

The estimate for the latter is difficult to understand. However, it can reasonably be expected that a positive χ_{\perp}^{nonloc} and the large value of χ_{av}^{nonloc} are not the only anomalies in the magnetic properties of cyclopropane. The in-plane components of carbon and proton shielding may also be biased by electron delocalization. This appears even more plausible for the huge in-plane shielding recently evaluated for a probe in the center of the cyclopropane ring.²⁴

A definite answer to these points can be obtained by quantum chemical calculations. In 1983 Lazzarotti and Zanasi²⁵ reported the first ab initio model of magnetically induced current density for cyclopropane in a field $\mathbf{B} = B_z\mathbf{e}_z$ orthogonal to the molecular plane, adopting a coupled Hartree–Fock (CHF) common origin (CO) procedure. Although that model does not provide a correct description of electron flow in either the vicinity of the carbon nuclei or about the midpoint of the C–C bonds, it showed two unexpected important features: (i) a central paratropic vortex and (ii) three local diatropic vortices circulating about the C nuclei.

As the model gave no clear evidence of a ring current, the authors attributed the large σ_{zz}^H value to the (upfield) diamagnetic shift caused by (ii). Subsequent ab initio calculations have unambiguously documented a delocalized current,^{26,24,27} although its contribution to the magnetic properties of cyclopropane has not yet been quantified.

An RCM of cyclopropane providing an interpretation of its seemingly anomalous magnetic susceptibility, reported by Bader and Keith,²⁶ is reviewed in Section 1 of the Supporting Information. A more recent model proposed by Fowler, Baker, and Lillington (FBL),¹⁹ consistent with the literature attribution of σ -aromaticity to C₃H₆,^{11,28,29,15,30–32} is critically revised by methods outlined in Section 4, where correct criteria for vortex–saddle merging are outlined, see also Section 2 of the Supporting Information.

Recent calculations were reported by Fliegl et al., who concluded that ring currents in C_3H_6 are not negligible because the total current strength (evaluated by numerical integration of the net current \mathbf{J}^B over a half-plane through the midpoint of a C–C bond and normal to it) is 10.0 nA/T, which is only 1.8 nA/T smaller than for benzene.²⁷ These authors claim that the strongest diatropic current flow appears in the molecular plane for both cyclopropane and benzene molecules.³³ Accordingly, allowing for the ring-current criterion of aromaticity, cyclopropane would be almost as aromatic as benzene, even though it is fully saturated.²⁷ The results of Fliegl et al.²⁷ are discussed in Section 6.

Allowing for the present state of affairs, this paper is primarily meant to complete the \mathbf{J}^B current density model of a previous article²⁴ at the Hartree–Fock level by (i) investigating the main features of the \mathbf{J}^B vector field for a magnetic field B_{\perp} parallel to a C_2 symmetry axis, (ii) analyzing the isolated critical points of \mathbf{J}^B and the related phase portraits, in terms of local effects, (iii) interpreting the difference between $3\chi_{av}^{nonloc}$ and $\Delta\chi^{nonloc}$, and (iv) rationalizing the different contributions to the components of magnetic tensors. Points (i) and (iii) will at once show that the huge average NICS of cyclopropane, sometimes considered an indicator of super σ -aromaticity,^{11,28,29,34,35} is instead an unreliable quantifier of magnetotropy. These results add further evidence against the idea that NICS might really be used to assess aromaticity, the σ -aromatic paradigm of cyclopropane in particular.

2. Quantum Mechanical Current Density Models

The most interesting characteristics of a \mathbf{J}^B field are observed in the proximity of its stagnation points (SP) at which the modulus $|\mathbf{J}^B|$ vanishes. In the neighborhood of an SP with position \mathbf{r}_0 the field is described by a truncated Taylor series,

$$\mathbf{J}_{\gamma}^B(\mathbf{r}) = (r_{\alpha} - r_{0\alpha})[\nabla_{\alpha}\mathbf{J}_{\gamma}^B]_{\mathbf{r}=\mathbf{r}_0} + \frac{1}{2}(r_{\alpha} - r_{0\alpha})(r_{\beta} - r_{0\beta})[\nabla_{\alpha}\nabla_{\beta}\mathbf{J}_{\gamma}^B]_{\mathbf{r}=\mathbf{r}_0} + \dots \quad (1)$$

Within the linear approximation, an exhaustive compilation of all possible phase portraits about an SP in three-dimensional flow has been reported by Reyn,³⁶ in connection with the canonical forms of the 3×3 Jacobian matrix $\nabla_{\alpha}\mathbf{J}_{\gamma}^B(\mathbf{r}_0)$. Since the local regime depends on the eigenvalues of this matrix, SPs are classified in terms of an Euler (*rank, signature*) label.^{26,37–43} The rank r is the number of nonvanishing eigenvalues of the Jacobian matrix and the signature s is the excess of eigenvalues with positive over negative real part. An SP is also classified in terms of its topological index ι .^{44–46} SPs of type $(3, \pm 1)$ are called *isolated*. The corresponding phase portrait is that of *saddle node*,³⁶ if all the eigenvalues are real, or a *focus*, if two eigenvalues are complex conjugate. In the latter case, the local trajectories spiral inward or outward in the direction of the third (real) eigenvalue.

Continuous, open or closed, paths of $(2,0)$ points are referred to as stagnation lines (SL), consisting of either *vortex*, also referred to as *center* points (index $\iota = +1$, two

nonvanishing complex conjugated eigenvalues), or *saddle* points (index $\iota = -1$, two nonvanishing real eigenvalues with the same magnitude and opposite sign). Some examples have previously been reported.^{24,47,26,48,41,49–53}

The three-dimensional structure of a current density vector field is described by a stagnation graph (SG), which collects all isolated SPs and SLs. A $(2,0)$ SL may branch out at $(0,0)$ critical points. The Gomes theorem provides an index conservation constraint, $\iota_0 = \sum_{k=1}^m \iota_k$, for an SL with index ι_0 splitting into m new lines at a branching point.^{38–40,54}

3. RCM of Cyclopropane in a Magnetic Field Perpendicular to the Plane of Carbon Nuclei

The SG for cyclopropane in a magnetic field B_{\parallel} perpendicular to the plane of the carbon nuclei has previously been constructed employing the method of continuous transformation of the origin of the current density-diamagnetic zero (CTOCD-DZ).²⁴ As all the CTOCD variants of the current density are invariant in a translation of the origin of the gauge,⁵⁶ this CTOCD SG is also invariant. It illustrates in a compact way the main features of the induced current density vector field with

$$D_{3h}(C_{3h}) \equiv \{E\ 2C_3\ 3TC_2\ \sigma_h\ 2S_3\ 3T\sigma_v\} \equiv \bar{6}m2$$

magnetic symmetry³⁷ (denoting time-inversion by T) observable in Figure 1: a central paratropic vortex, i.e., a whirlpool circulating about the red SL coincident with the C_3 symmetry axis and parallel to B_{\parallel} , three diatropic vortices, sustained by the carbon sp^3 hybrid orbitals forming the C–H bonds, and a peripheral flow of “ring currents”, with the shape of a topological torus. Each of the three green SLs passes through a C nucleus and both C and H nuclei sit inside the diatropic vortices flowing around. The SGs obtained via other calculations of increasing accuracy are reported as Supporting Information to document convergence to the Hartree–Fock limit.

A major difference from other compounds usually regarded as π -aromatic on the magnetic criterion⁴⁸ is the absence of a spatial vortex about the center of C–C bonds,²⁶ the green SL of a diatropic vortex being replaced by a blue saddle SL extending up to a pair of branching points distant more than 10 bohr from the center of the molecule.^{24,26} Each saddle SL passes between two green isolated $(3, \pm 1)$ SPs on the plane of the C nuclei, at a short distance from one another. Two more pairs of green SPs, symmetrically placed above and below the molecular plane, are also observed in the region of each C–C bond. All of them belong to a set of foci (SPs characterized by one real and two complex conjugated eigenvalues of the Jacobian matrix, see Section 2). They are connected by diatropic streamlines spiralling about them, above and below the σ_h symmetry plane. Asymptotic lines connect the 18 foci to the 6 saddle nodes, see Figure 2.

Since crossing of the σ_h plane is forbidden by magnetic symmetry,³⁷ the trajectories in the neighborhood of the foci about the midpoint of C–C bonds are *limit cycles*, which may incorrectly be interpreted as vortices in current density maps on the molecular plane, see Figure 1 of the Supporting

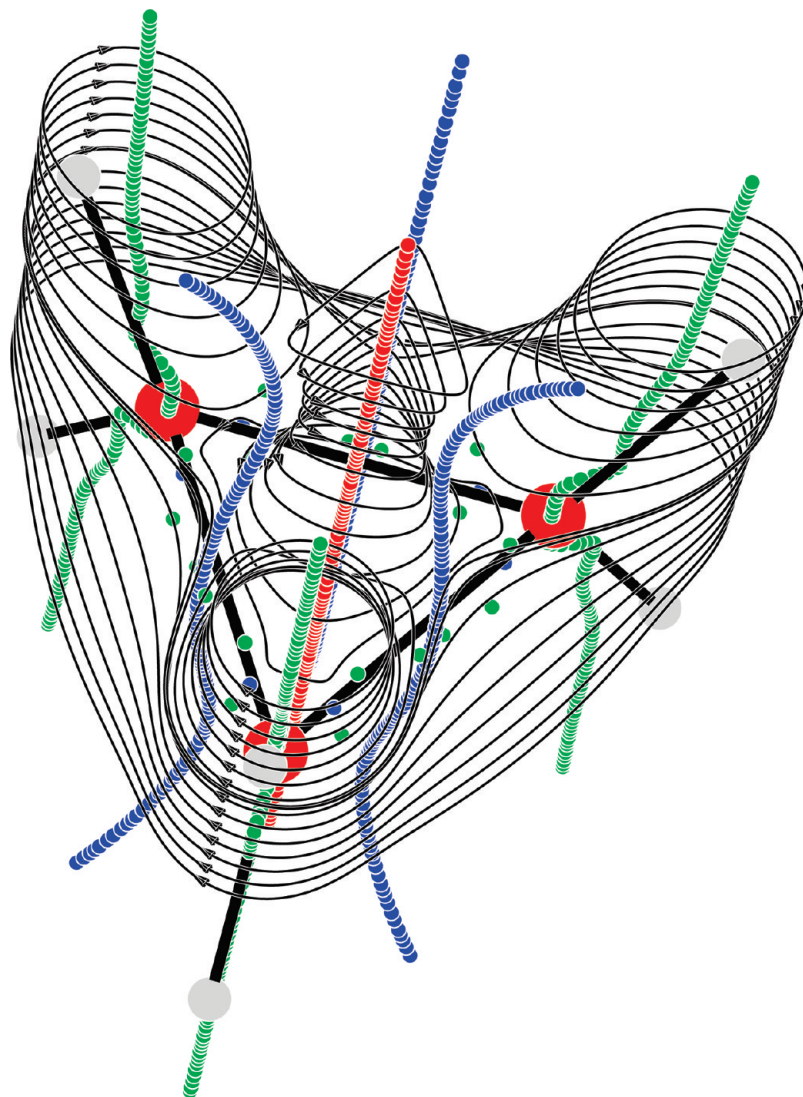


Figure 1. Spatial ring-current model of cyclopropane in a magnetic field perpendicular to the plane of the carbon nuclei, superimposed to the stagnation graph. A paratropic axial vortex circulates about the C_3 symmetry axis. Diatropic vortices, sustained by the carbon sp^3 hybrid orbitals forming the C–H bonds, are embedded within a delocalized peripheral flow (which can be described as a σ -ring current), having higher intensity $|\mathbf{J}^B|$ in a domain with the shape of a topological torus. $|\mathbf{J}^B|$ goes smoothly to 0 at ∞ . The modulus of the current vanishes along the stagnation lines in the central region of each vortex. These continuous open lines are represented by a sequence of green (red) dots for diatropic (paratropic) regime. Saddle stagnation lines crossing the σ_h plane slightly outside the midpoint of a C–C bond are blue. Isolated $(3, \pm 1)$ foci (saddle nodes) are indicated in green (blue).⁵⁵

Information. Low-resolution current density maps displaying only the planar regime on σ_h are insufficient to visualize the local regime. The analysis of the eigenvalues of the Jacobian matrix, eq 1, and a blow-up of this figure are required to reveal the presence of the foci, see the Supporting Information of a previous paper.²⁴

One could hardly find a better proof that *spatial* RCMs, illustrated in Figures 1 and 2, are needed to understand the magnetic response of cyclopropane.⁵⁵

4. Merging of Vortex and Saddle Flow in Cyclopropane

FBL proposed a model merging three diatropic bond vortices to interpret the diatropic ring current and the central paratropic vortex observed in their current density maps for cyclopropane.¹⁹ In those maps important patterns were not

shown, e.g., the strong C-centered circulations and the fine structure of the current density field \mathbf{J}^B , whose most striking feature is probably the saddle—rather than vortex—character of the stagnation points close to the midpoints of the C–C bonds, as already reported in previous papers.^{26,24} Thus a different model is required to get a qualitatively correct representation of \mathbf{J}^B of cyclopropane, allowing for the stagnation graph of Figure 1 and the pseudostagnation graph of Figure 3. In the latter, the isolated $(3, \pm 1)$ SPs are connected by continuous paths of points at which the out-of-plane $\mathbf{J}_z^B \neq 0$, whereas the modulus of the in-plane current \mathbf{J}_{xy}^B vanishes. For that reason these paths are called pseudostagnation lines.

Starting from examination of \mathbf{J}^B far from the molecular plane in Figure 1, the set of one central SL and six pseudostagnation lines can be interpreted in terms of a basic

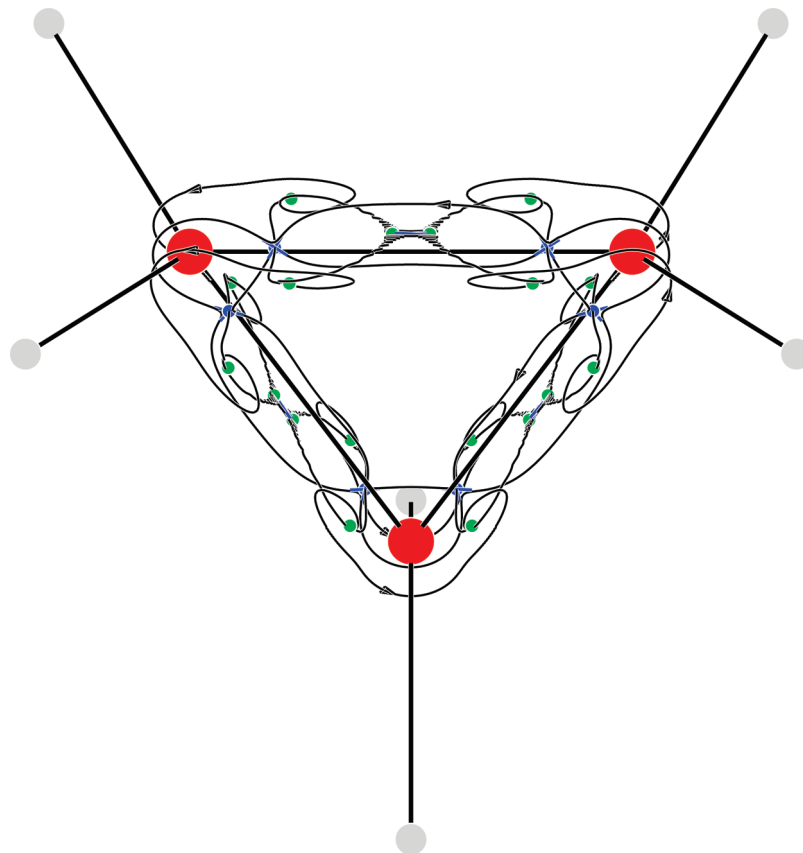


Figure 2. Spiral flow connecting the 18 isolated $(3, \pm 1)$ foci (in green) and the 6 isolated $(3, \pm 1)$ saddle nodes (in blue, marked by a cross) in the C–C bond regions of cyclopropane in a magnetic field B_{\parallel} .

center–saddle–center pattern roughly above a C–C bond; that pattern is consistent with the summation of two separated diatropic C-centered vortices.⁵⁷ Nonetheless, further investigation requires the analysis of the isolated SPs.

The 6 foci and 6 saddle nodes on the C-plane as well as the 12 out-of-plane foci have a three-dimensional nature; as all of the eigenvalues of their Jacobian are nonzero, they should be analyzed in terms of genuine three-dimensional fields, which excludes the possibility of taking sums of purely rotational fields as in a previous paper.⁵⁷ However, an analogous analysis is possible by formally setting $J_z^{\mathbf{B}} = 0$ and considering only the field over the xy plane, $\mathbf{J}_{xy}^{\mathbf{B}} = J_x^{\mathbf{B}}\mathbf{e}_1 + J_y^{\mathbf{B}}\mathbf{e}_2$, which is sufficient to determine the magnetic properties, although it has nonzero divergence due to the cancellation of the $J_z^{\mathbf{B}}$ component.

The isolated $(3, \pm 1)$ SPs and the $(2, 0)$ saddle SLs of the $\mathbf{J}^{\mathbf{B}}$ field are also present in the $\mathbf{J}_{xy}^{\mathbf{B}}$ field. However, in the pseudostagnation graph for the latter in Figure 3, the isolated single points are connected by six closed pseudostagnation loops, symmetrically placed on both sides of each true saddle line intersecting σ_h about the midpoint of a C–C bond, see also Figure 1. The pseudostagnation loops lie inside domains bounded by asymptotic streamlines shown in Figure 2. In the following we will assume, without loss of accuracy, that the $(3, \pm 1)$ saddle nodes (foci) can be treated as planar saddles (centers), which is consistent with the fact that the small local $J_z^{\mathbf{B}}$ component is neglected. We have accordingly found that each of the three pairs of stagnation loops of $\mathbf{J}_{xy}^{\mathbf{B}}$ in Figure 3 can be interpreted in terms of three diatropic

circulations: two of them being centered on the C atoms and one on the C–C bond.

To obtain this result, we have considered the summation of three purely rotational two-dimensional diatropic vortices with (x, y) centers in $(R_1, 0)$, $(R_2, 0)$ and $(R_3, 0)$ and the general expression $\mathbf{J}^{\mathbf{B}(i)} = J^{\mathbf{B}(i)}(r_i)\hat{\theta}_i$, where the $J^{\mathbf{B}(i)}(r_i)$ functions vanish only at 0 and ∞ , $i = 1, 2$, and 3, and r_i and θ_i are the polar coordinates with respect to the i -th center. The three components $J_x^{\mathbf{B}(i)}$ have the same sign above and below the x direction, where they vanish. Therefore, stagnation points can only be found on the x axis if the three components $J_y^{\mathbf{B}(i)}$ cancel out. Assuming real (imaginary) eigenvalues for the two-dimensional Jacobian leads to the conclusion that a stagnation point will be a saddle (vortex) if

$$\frac{\partial |J_{\theta}^{\mathbf{B}(1)}|}{\partial r_1} - \frac{\partial |J_{\theta}^{\mathbf{B}(2)}|}{\partial r_1} - \frac{\partial |J_{\theta}^{\mathbf{B}(3)}|}{\partial r_1}$$

is negative (positive). This condition reduces to that previously analyzed for two circulations if $|J_{\theta}^{\mathbf{B}(2)}| = 0$.⁵⁷

The shape of the $J_{\theta}^{\mathbf{B}(i)}$ functions is clearly a crucial factor for determining the overall pattern. The ab initio current density can in general be expressed as a sum of Gaussian terms.⁵⁷ However, for the sake of simplicity, we will make use here of a single exponential form. The form chosen is $J_{\theta}^{\mathbf{B}(i)} = I_i N_i r_i^2 \exp(-a_i r_i^2)$, where the shape of the function is determined by two parameters, a_i and I_i , and the normalization constant is $N_i = 1/\int_0^{\infty} r_i^2 \exp(-a_i r_i^2) dr_i$. We have first considered the summation of three identical circulations of

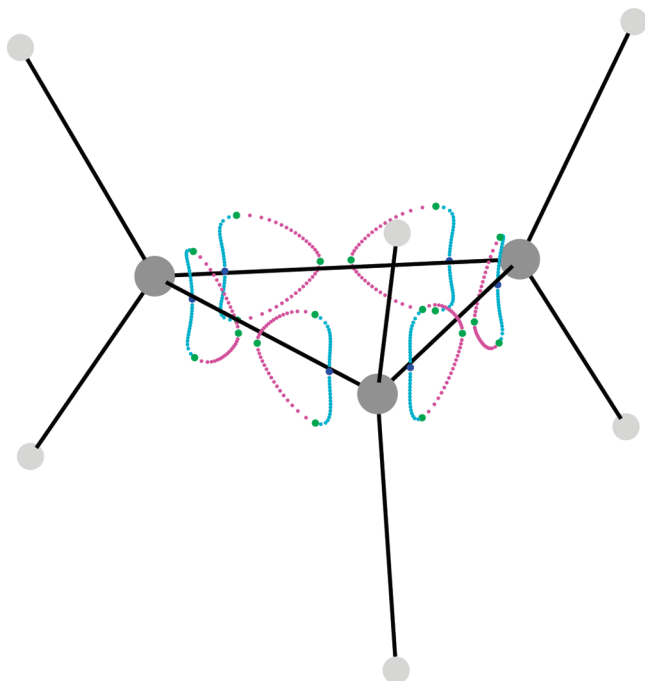


Figure 3. Pseudostagnation graph of the planar \mathbf{J}_{xy}^B field. The closed loops are continuous paths of points at which the modulus $|\mathbf{J}_{xy}^B| = 0$. They contain $(3, \pm 1)$ isolated points, observed also in Figures 1 and 2, where the total current density $|\mathbf{J}^B|$ vanishes. Green (blue) dots correspond to foci (saddle nodes). Magenta (cyan) paths indicate that spiral (saddle) flow is observed on \mathbf{J}_{xy}^B cross-sectional streamline plots parallel to σ_h . The seven SLs characterizing the \mathbf{J}^B field, see Figure 1, have been omitted for clarity.

unitary intensity I placed at $-0.75, 0$, and 0.75 \AA on the x axis. Figure 4 shows that, on changing the common a parameter, the number and type of the stagnation points changes. Starting from a single center for very broad functions (small a) and sharpening the $J_{\theta}^{B(i)}$ functions, one gets three stagnation points, corresponding to two circulations separated by a saddle. For still sharper functions, seven stagnation points appear: four circulations separated by three saddles, as found on the molecular plane of cyclopropane.

Setting now $a_1 = a_2 = a_3 = 10 \text{ \AA}^{-2}$, a value at which the 7 stagnation points occur, we have investigated the behavior of the pseudostagnation lines on moving further from the molecular plane. To simulate the fact that C-centered circulations, sustained by the C–H electron density, extend higher than the bond-centered circulations, we varied the intensity of the central flow. Figure 5 displays the position and the nature of the stagnation points as the intensity I_2 is changed from 0 to $2I_1$. The vertical axis has been reversed to make easier the comparison with Figure 3. It can be seen that, for a low intensity of the central circulation, the current pattern is substantially that expected for only two separated homotopic circulations.

However when the strength of the central flow is comparable with that of the circulations nearby, the pattern of seven stagnation point starts to appear. Upon further increase, the central circulation dominates the current pattern, although the presence of the circulations at its sides can still be predicted from the three center–saddle–center stagnation

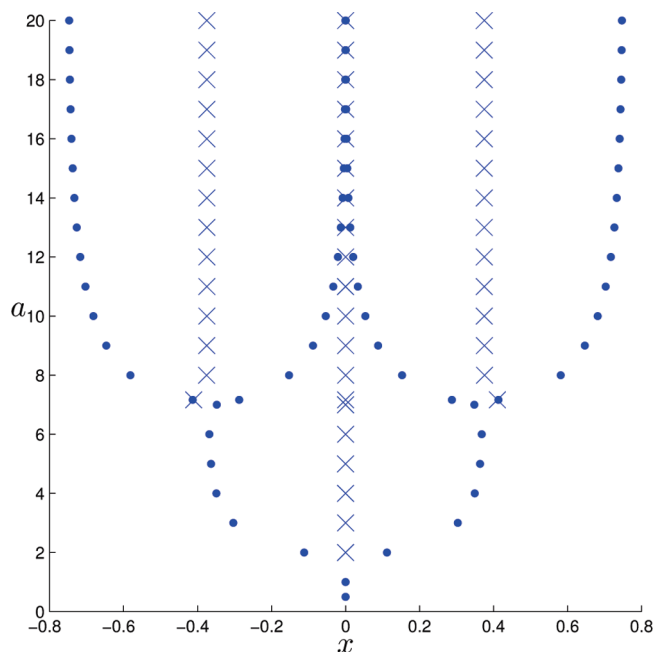


Figure 4. Nature and location of the stagnation points of the \mathbf{J}_{xy}^B field obtained by summing three collinear diatropic vortices of equal shape and intensity, placed at $x = -0.75, 0$, and 0.75 \AA . The computations were repeated for different values of the exponent a (see text). Centers and saddles are indicated by \cdot and \times , respectively.

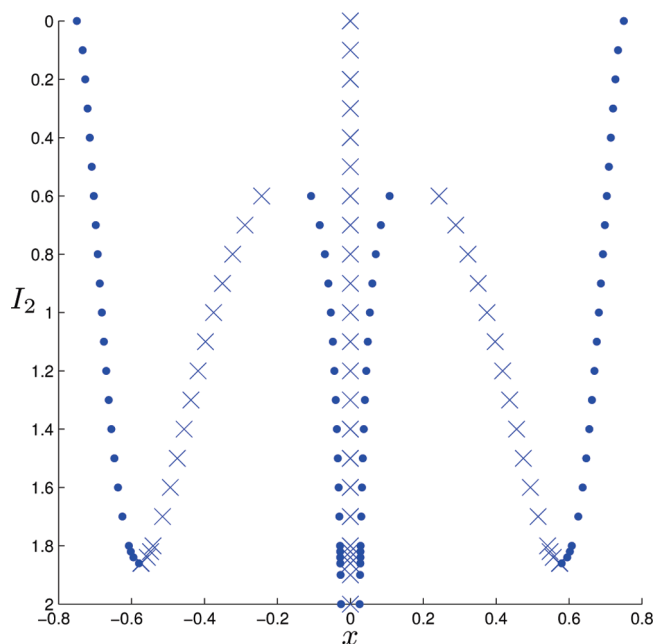


Figure 5. Nature and location of the stagnation points of the \mathbf{J}_{xy}^B field obtained by summing three collinear diatropic vortices located, as in Figure 4, and with $a = 10 \text{ \AA}^{-2}$. The computations were repeated for different values of the I_2/I_1 ratio. Centers and saddles are indicated by \cdot and \times , respectively.

points close to $x = 0$, rather than a single center. The upper part of Figure 5, with the disappearance of four stagnation points, nicely parallels the pseudostagnation graph of the \mathbf{J}_{xy}^B field displayed in Figure 3.

5. Magnetic Response Tensors of Cyclopropane

Attempts are usually made to infer molecular current density models by experimental values of magnetic susceptibility¹² and NMR chemical shift⁵⁸ as well as nonmeasurable NICS.^{13,14} However, it is not possible, in general, to construct a plausible global model of $\mathbf{J}^{\mathbf{B}}$ field only allowing for a few numbers. Just the other way around, one can reasonably argue that a *falsifiable* current density model should be developed in advance. After any possible tests on its ability to rationalize all available data, such a model is accepted or rejected. Improved versions can later be sought, still requiring their falsifiability.

Therefore, we find it expedient to start by defining the magnetic properties of a molecule in terms of the quantum mechanical induced current density. When the *subobservable*⁵⁹ $\mathbf{J}^{\mathbf{B}}$ is available, one can treat it as a completely classical quantity, forgetting about the quantum mechanical procedure used to obtain it and rely on the law of classical electrodynamics for the interpretation of magnetic response.

Denoting by $\epsilon_{\alpha\beta\gamma}$ the Levi–Civita unit tensor and using the implicit summation rule for repeated suffixes according to tensor notation, the orbital magnetic dipole moment induced in the n electrons of a molecule by an external magnetic field with flux density \mathbf{B} is evaluated by the Ampere law assuming linear response:

$$\Delta\langle m_{\alpha} \rangle = \chi_{\alpha\beta} B_{\beta} = -\frac{1}{2} \epsilon_{\alpha\beta\gamma} \int J_{\beta}^{\mathbf{B}}(\mathbf{r}) r_{\gamma} d^3r \quad (2)$$

The magnetic field induced at an observation point \mathbf{R} is determined by the Biot–Savart law:

$$\Delta\langle B_{\alpha}^I(\mathbf{R}) \rangle = -\sigma_{\alpha\beta}(\mathbf{R}) B_{\beta} = \frac{\mu_0}{4\pi} \epsilon_{\alpha\beta\gamma} \int J_{\beta}^{\mathbf{B}}(\mathbf{r}) \frac{R_{\gamma} - r_{\gamma}}{|\mathbf{R} - \mathbf{r}|^3} d^3r \quad (3)$$

Introducing the current density tensor⁶⁰ via the derivative:

$$\mathcal{J}_{\alpha}^{B_{\beta}}(\mathbf{r}) = \frac{\partial}{\partial B_{\beta}} J_{\alpha}^{\mathbf{B}}(\mathbf{r}) \quad (4)$$

the magnetizability tensor is evaluated by

$$\chi_{\alpha\delta} = \frac{1}{2} \epsilon_{\alpha\beta\gamma} \int r_{\beta} \mathcal{J}_{\gamma}^{B_{\delta}}(\mathbf{r}) d^3r \quad (5)$$

and the shielding tensor at \mathbf{R} is obtained as

$$\sigma_{\alpha\delta}(\mathbf{R}) = -\frac{\mu_0}{4\pi} \epsilon_{\alpha\beta\gamma} \int \frac{r_{\beta} - R_{\beta}}{|\mathbf{r} - \mathbf{R}|^3} \mathcal{J}_{\gamma}^{B_{\delta}}(\mathbf{r}) d^3r \quad (6)$$

If \mathbf{R} coincides with the position \mathbf{R}_I of the I -th nucleus, carrying an intrinsic magnetic dipole $m_{I\alpha}$, the quantity $\sigma_{\alpha\beta}(\mathbf{R}_I) \equiv \sigma_{\alpha\beta}^I$ defines the magnetic shielding tensor of that nucleus. The integrand function is interpreted as a shielding density second-rank tensor,^{61,62} for instance

$$\Sigma_{zz}^I(\mathbf{r}) = -\frac{\mu_0}{4\pi} \epsilon_{z\beta\gamma} \frac{r_{\beta} - R_{I\beta}}{|\mathbf{r} - \mathbf{R}_I|^3} \mathcal{J}_{\gamma}^{B_z}(\mathbf{r}) \quad (7)$$

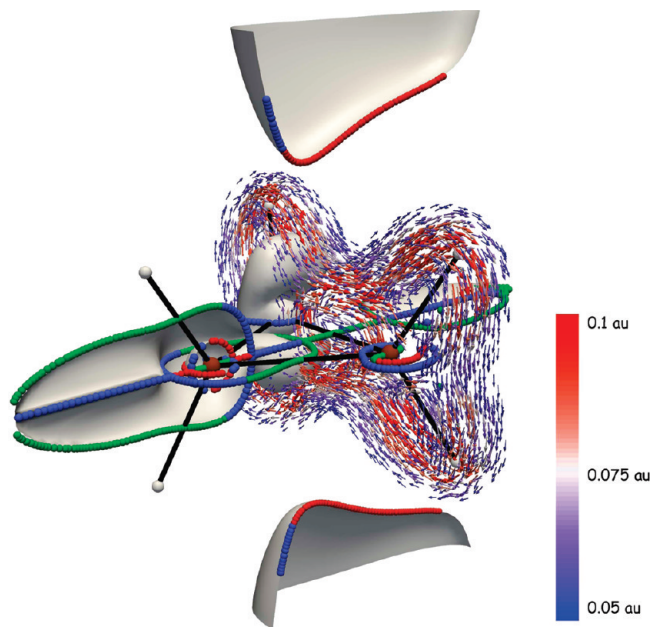


Figure 6. The strong “ring current” flowing around the $\text{CH}_2\text{—CH}_2$ fragment of the cyclopropane molecule in a magnetic field of 1 au, applied along the $C_2 \equiv x$ symmetry axis. Only current densities with $|\mathbf{J}^{\mathbf{B}}|$ between 0.05 (blue arrows) and 0.1 au (red arrows) are plotted. The maximum modulus is ≈ 0.11 au. The gray surfaces represent isoshielding density regions $\Sigma_{zz}^I(\mathbf{r}) = 0.0$ au, for any ghost atom I along the C_2 axis. Vortical and saddle stagnation lines lie on these surfaces, see also Figure 7.

is the zz component of the shielding density for nucleus I at a point \mathbf{r} in a given domain.

This function is usually plotted over a plane, as in Figures 11 and 12, to analyze shielding/deshielding mechanisms operating in different basins of the current density field via a few prescriptions illustrating the effect of $\mathbf{J}^{\mathbf{B}}$ at point \mathbf{r} on $\Sigma_{zz}^I(\mathbf{r})$.⁴⁹ Isoshielding density surfaces can also be visualized, see Figures 7 and 8. Although *all* the planes perpendicular to a given direction, e.g., z , provide an infinitesimal “slice” contribution to the total induced field (eq 6), in practice one does not need to examine the density (eq 7) over a large number of plot planes. Usually only a few are taken into account, those from which sizable contributions are expected to arise, e.g., planes of nearly maximum charge distribution can be sampled.

It is important to recall that the induced orbital moment (eq 2) and the magnetic susceptibility (eq 5) are *global* properties, proportional to the area enclosed by a wide domain of induced current loops,⁶³ whereas the magnetic shielding (eq 6) at \mathbf{R} is mainly determined by the flow in a small region about the probe, as it depends on the second inverse power of the distance $|\mathbf{r} - \mathbf{R}|$ from the observation point. Therefore the components of the magnetic tensors $\chi_{\alpha\beta}$, $\sigma_{\alpha\beta}^I$, and possibly NICS, provide different, complementary pieces of information.

Near Hartree–Fock CTOCD estimates of the magnetic susceptibility $\chi_{\alpha\beta}$ and nuclear shieldings $\sigma_{\alpha\beta}^I$ of cyclopropane have recently been reported using an extended (13s10p5d2f/8s4p1d) basis set containing 435 primitive Gaussian functions. Calculations were carried out by the SYSMO computer

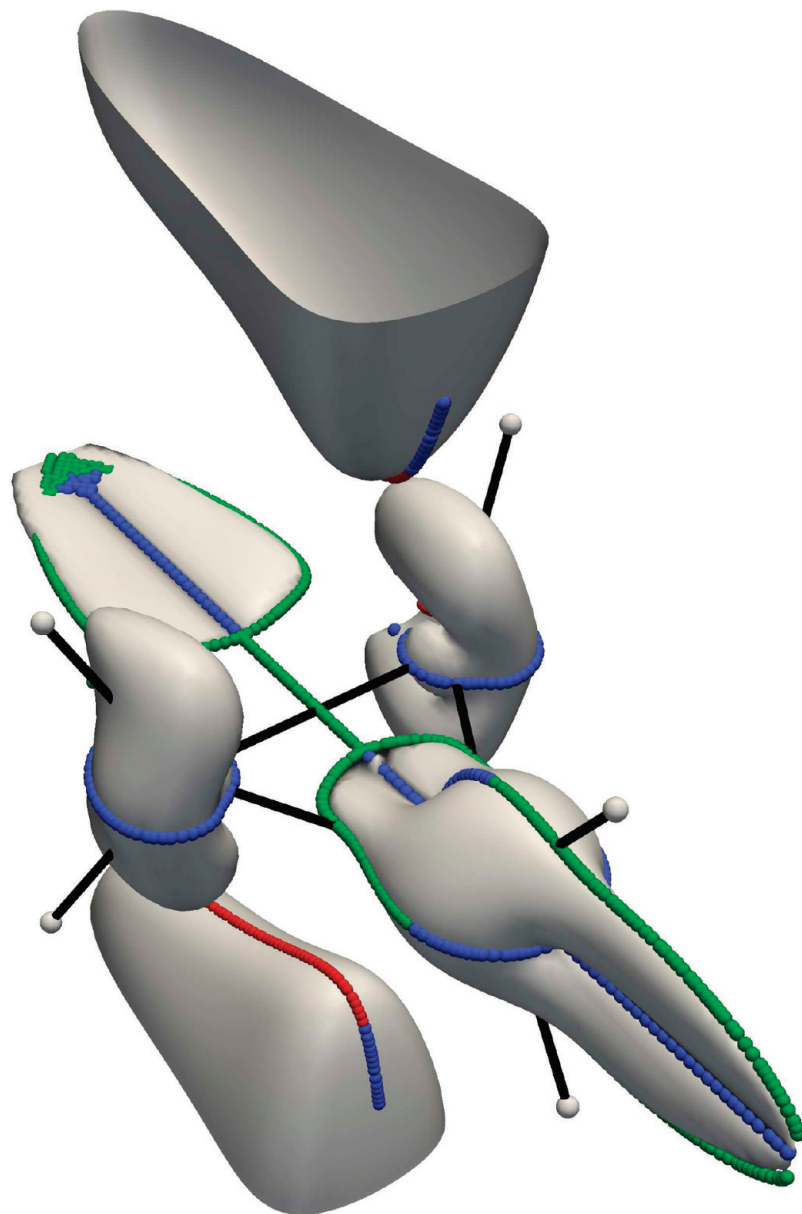


Figure 7. The stagnation graph of cyclopropane in a magnetic field parallel to the $C_2 \equiv x$ symmetry axis. The stagnation lines lie on isoshielding density surfaces $\Sigma'_{xx}(\mathbf{r}) = 0.0$ au, represented in gray, for any dummy atom l along C_2 .

code.⁶⁴ The high quality of the calculations was assessed by a number of criteria.²⁴ It should be emphasized that the magnetic tensors of cyclopropane calculated by CTOCD procedures are invariant in a gauge translation.^{60,65,37} The RCM developed in the present paper is required to explain sign and magnitude of the diagonal components of these tensors and to elucidate the source of their strong anisotropy by applying a few simple rules outlined hereafter.

According to eqs 2–6, the electronic magnetic moment $\Delta\langle m_\alpha \rangle$ and the magnetic field $\Delta\langle B'_z(\mathbf{R}) \rangle$ induced at position \mathbf{R} by an external field B_z , the magnetizability component χ_{zz} , and the nuclear shielding component $\sigma_{zz}(\mathbf{R})$ are determined only by the components J'_x and J'_y of the current density in the xy plane. The paramagnetic component J'_z has no effect on χ_{zz} and $\sigma_{zz}(\mathbf{R})$. These statements are valid for cyclic permutations of x , y , and z .

The diatropic electronic ring currents flowing in planes parallel to the $\sigma_h \equiv \sigma_{xy}$ plane of a conjugated cyclic molecule

in the presence of a magnetic field B_z : (i) exalt the out-of-plane component χ_{zz} of the magnetic susceptibility, and consequently increase the magnitude of the anisotropy $\Delta\chi$ and the average magnetic susceptibility χ_{av} and (ii) enhance (diminish) the out-of-plane component σ'_{zz} of a real or dummy nucleus l , placed at \mathbf{R}_l , inside (outside) the ring. In correspondence with the increase (decrease) of σ'_{zz} , one observes an upfield (downfield)—also called diamagnetic (paramagnetic)—contribution to the NMR chemical shift from a reference compound, usually tetramethylsilane (TMS), $\delta^l = \sigma_{av}^{ref} - \sigma_{av}^l$. By reversing the direction of \mathbf{J}^B , i.e., for a paratropic current, the sign of the contributions mentioned above is also reversed.

A familiar example widely discussed in the literature is benzene, whose π -ring currents determine observable effects: high anisotropy $\Delta\chi$, big χ_{av} , and downfield chemical shift of protons as well as a big positive value of the average

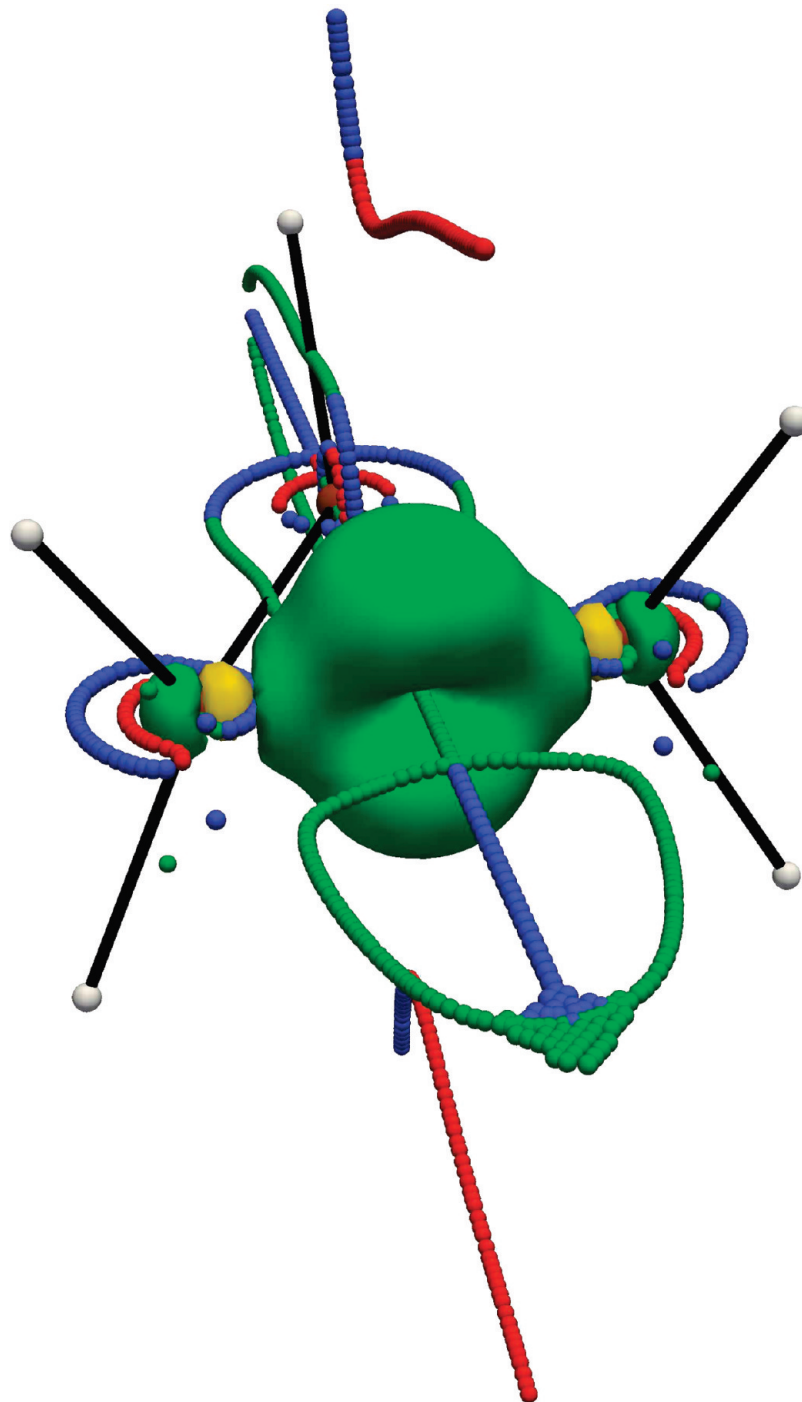


Figure 8. Isoshielding density surface $|\Sigma_{xx}^I(\mathbf{r})| = 0.05$ au, for a dummy atom I at the midpoint of a C–C bond. Green (yellow) portions denote positive diamagnetic (negative paramagnetic) contributions.

central shielding, or a big negative average NICS, which is defined as the same quantity with negative sign.¹³

However, it should be emphasized that the ring currents can only bias the out-of-plane component of the magnetic tensors of conjugated planar (poly)cyclic molecules. Properties observed in systems in disordered phase account for only one-third of the effect arising from the ring currents. Sizable, or sometimes overwhelming contributions from mechanisms other than ring currents, affecting the in-plane components of the magnetic tensors, should carefully be examined, and any theoretical assessment of

the strength of delocalized ring currents in planar conjugated carbon cycles should take only the out-of-plane components χ_{zz} and σ_{zz}^I into account.⁶⁶

For a ghost atom I , with coordinate \mathbf{R}_I , placed anywhere along the $C_3 \equiv z$ symmetry axis of C_3H_6 , the equation $\Sigma_{zz}^I(\mathbf{r}) = 0$ defines a surface of points \mathbf{r} of vanishing shielding density. The same statement holds for $\Sigma_{xx}^I(\mathbf{r}) = 0$, with I a dummy atom along the $C_2 \equiv x$ symmetry axis. Stagnation lines lie on the zero-isoshielding surfaces, as shown in Figure 6 by superimposing the stagnation graph to some portions of the $\Sigma_{xx}^I(\mathbf{r}) = 0$ domain. Another view of this pattern is

observed in Figure 7 for all dummy atoms I lying on C_2 , with arbitrary \mathbf{R}_I .

Abnormally high calculated values of average NICS¹³ are considered to be consistent with σ -diatropicity of C_3H_6 .^{28,29,34,35} In fact, cyclopropane provides crucial evidence on the failure of the average NICS¹³ as a measure of diatropicity.²⁴ The calculated NICS $\equiv -\sigma_{av}^{CM}$ is as big as -44.9 ppm, a value comparable with the GIAO NICS RHF/6-311+G(d,p) estimate -43.0 given by Sauers,³⁵ and with similar values by others.^{28,34} Noticeably enough, σ_{av}^{CM} is dominated by the enormous in-plane component $\sigma_{\perp}^{CM} = 50.9$ ppm, a case predicted in a previous paper.⁶⁶ The out-of-plane σ_{\parallel}^{CM} component is ≈ 18 ppm smaller. Therefore, neither the average NICS¹³ nor NICS _{\parallel} ^{66,67} can be used as magnetic aromaticity quantifiers in cyclopropane.

This peculiarity of C_3H_6 is explained by the RCM developed here. In a field B_{\perp} applied in the direction of a C_2 symmetry axis, intense delocalized currents (with maximum modulus ≈ 0.11 au for an applied magnetic field of 1 au) take place in the bent “banana bond” lying outside of the direction interconnecting two carbon nuclei,⁶⁸ and about the C–H bonds, as shown by Figure 6 in the text and Figure 6 in the Supporting Information. It is this current, with strength 15.7 nA/T (evaluated as a flux integral over a suitably defined “bond basin” in Section 6), flowing around the CH_2-CH_2 and methylene moieties that determines the huge value of the σ_{\perp}^{CM} .

Quite remarkably, a similar pattern was observed for ethylene, another noncyclic system which sustains an annular weaker current (with modulus 0.075 au, for $|\mathbf{B}| = 1$ au) on the molecular plane, see Figure 12 of a previous article.⁶⁹ The SG of the ethylene molecule in a magnetic field perpendicular to σ_h , see Figure 6 of the same paper⁶⁹ contains loops similar to those observed in the vicinity of the C nuclei in Figure 6.

It may be useful to complete the analysis of the different contributions to the central shielding (and NICS) values calculated for cyclopropane by comparison with corresponding data for benzene. Allowing for a calculation adopting the same basis set for the sake of consistency,⁴⁸ we obtained $\sigma_{\perp}^{CM} = 5.38$, $\sigma_{\parallel}^{CM} = 18.40$, and $\sigma_{av}^{CM} = 9.72$ ppm for C_6H_6 . The σ and π contributions to the out-of-plane component, determined by conflicting mechanisms, are -18.86 and 37.26 , respectively (all values in ppm). Therefore, even if the $\sigma_{\parallel}^{CM} = 32.69$ ppm is smaller than σ_{\perp}^{CM} for cyclopropane, it is far larger than that for benzene and only 4.57 ppm smaller than the dominant π -contribution to benzene’s σ_{\parallel}^{CM} .

Recalling that the shielding density of a probe depends on the second inverse power of its distance from an element of current, according to the Biot–Savart law (eq 3), we conclude that total calculated σ_{\parallel}^{CM} values for the two molecules should be rationalized in terms of: (i) comparable values of $|\mathbf{J}^B|$, but (ii) different spatial extensions of the integration domains, and (iii) competing σ - and π -electron flow in benzene, which are absent in cyclopropane.

These results indicate that any conclusion on relative magnetic aromaticity in benzene and cyclopropane based on σ_{\parallel}^{CM} (and NICS _{\parallel}) values would not make sense.

6. Integrated Current Densities

The divergence of the stationary current density vector vanishes everywhere. Then, according to the Gauss theorem, also the flux Φ of the \mathbf{J}^B field vanishes for any closed surface S enclosing the volume V :

$$\Phi = \int_S \mathbf{J}^B \cdot d\mathbf{s} = \int_V \nabla \cdot \mathbf{J}^B dv = 0 \quad (8)$$

From this relationship, it is easy to show (see a forthcoming paper on the π character of ring current in aromatics)³³ that the integral of the current density crossing *any* arbitrarily chosen plane P bisecting the molecular domain vanishes:

$$\int_P \mathbf{J}^B \cdot d\mathbf{p} = 0 \quad (9)$$

In actual calculations, $\nabla \cdot \mathbf{J}^B$ is not zero, and eq 9 is not exactly fulfilled, except for symmetry reasons, e.g., for all planes containing an n -even symmetry axis C_n parallel to the inducing magnetic field \mathbf{B} , and for a symmetry plane perpendicular to \mathbf{B} . In general, the magnitude of the integral in eq 9 approaches zero on improving the quality of the calculation.

However, the integral in eq 9 is different from zero when P is a bounded portion of a plane, thus a measure of the strength of delocalized currents is assumed to be given by the flux of the \mathbf{J}^B vector through a suitably selected planar domain. For symmetric cyclic hydrocarbons, the halfplane bounded by the symmetry axis and bisecting a CC bond is a convenient choice of P to evaluate current strengths, also referred to as bond current susceptibilities.^{27,70–72}

In the case of benzene and other π -diatropic systems in a magnetic field \mathbf{B} at right angles to the molecular plane, *all* the streamlines of the π -current enter in the same direction the domain P of the integral $\int_P \mathbf{J}^B \cdot d\mathbf{p}$ defining the current strength. The largest $|\mathbf{J}^B|$ modulus values of the π -current of benzene are observed inside two toroidal regions of higher π -electron density, above and below the σ_h plane.⁷⁰ The π -currents flowing inside these Farnum–Wilcox tori⁷³ give an overwhelming contribution to the total current strength $\int_P \mathbf{J}^B \cdot d\mathbf{p}$.

On the other hand, the σ currents of most conjugated cyclic molecules, including benzene, are diatropic (paratropic) outside (inside) the ring.⁴⁸ Moreover, disconnected domains with the same regime frequently occur, which makes the definition of P more complicated. This is the case of cyclopropane, as observed for B_{\parallel} in Figure 9, which shows the cross-section of the induced current density modulus on a plane bisecting a C–C bond. The calculated current strength is 10.2 nA/T, the diatropic (paratropic) contribution being 11.6 (-1.4) nA/T. These estimates virtually coincide with those of a previous paper.²⁷

Figure 9 documents the effect of an intense diatropic circulation, with the shape of a topological torus, embedding the entire molecule. Such a ring current is sustained by delocalized σ -electrons shared by the three carbon atoms, as claimed by Dale Poulter et al.,⁷ and also by the hydrogen atoms. Therefore, one should not limit himself to consider the effect of electrons precessing in a *circle* which circumscribes the ring.⁷ The magnetic response of cyclopropane to

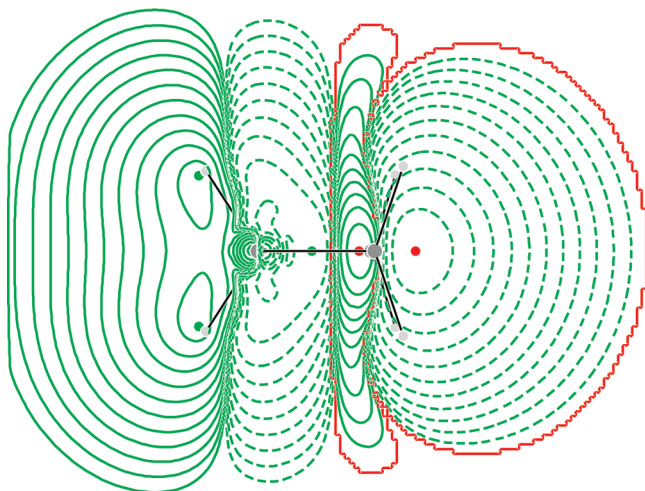


Figure 9. Contour plot of the cross-section of the current density modulus $|\mathbf{J}^{\mathbf{B}}|$ on a 14×14 bohr region of the xz plane spanned by the $C_2 \equiv x$ and the $C_3 \equiv z$ symmetry axes, bisecting a C–C bond. The applied magnetic field B_z , of magnitude 1 au, is orthogonal to the $\sigma_h \equiv xy$ symmetry plane. The H and C nuclei are represented in two shades of gray. Solid and dashed lines denote flow in opposite directions. Within the integration domain P , bounded by the red frontier, paratropic (diatropic) currents correspond to solid (dashed) contours. Red and green dots indicate extreme values of $|\mathbf{J}^{\mathbf{B}}|$ on the plot plane, $|\mathbf{J}^{\mathbf{B}}|_{\max}^{\rho} = 0.054$ ($|\mathbf{J}^{\mathbf{B}}|_{\max}^{\sigma} = 0.128$) au for the internal paratropic (external diatropic) flow. Contour values decrease in steps of $|\mathbf{J}^{\mathbf{B}}|_{\max}/2n$, for $n = 1, 2, \dots$. The total current strength, defined as the flux integral $\int_P \mathbf{J}^{\mathbf{B}} \cdot d\mathbf{p}$ from the domain within the red boundary, is 10.235 nA/T. The contributions from external (diatropic) and internal (paratropic) flow are 11.619 and -1.384 nA/T, respectively. Here, and in Figure 10, calculated current strengths do not show significant changes on enlarging the integration domain.

a magnetic field orthogonal to the carbon plane should more properly be interpreted in terms of a delocalized *toroidal* current flowing around carbon and hydrogen atoms.

The calculated current strength for $B_{\perp} = 1$ au, see Figure 10, is 15.7 nA/T, i.e., ≈ 1.5 times higher than that calculated for B_{\parallel} . The huge value of the $\sigma_{\perp}^{\text{CM}} = 50.9$ ppm is biased to a great extent by this peculiar “delocalized current without a carbon ring”, see also Figure 6 in the Supporting Information.

7. Rationalization of Magnetic Shieldings in Cyclopropane via Spatial RCMs

According to Fliegl et al., it is the ring strain that makes the cyclopropane electrons mobile, resulting in a strong magnetic ring current, which could be called “ring-strain current”, affecting magnetic shieldings in the same way as ring currents do.²⁷ In fact, a ring current flowing in an annular region of the plane of the C nuclei, around the methylene groups, and splitting into two branches along the C–H bonds of cyclopropane in a magnetic field B_z normal to the molecular plane can be displayed, see Figure 3 of a previous paper,²⁴ by selecting trajectories with a modulus varying between 0.05 and 0.1 au, for B_z with strength 1 au, and by cutting currents with much higher intensities in the proximity of the C atoms.

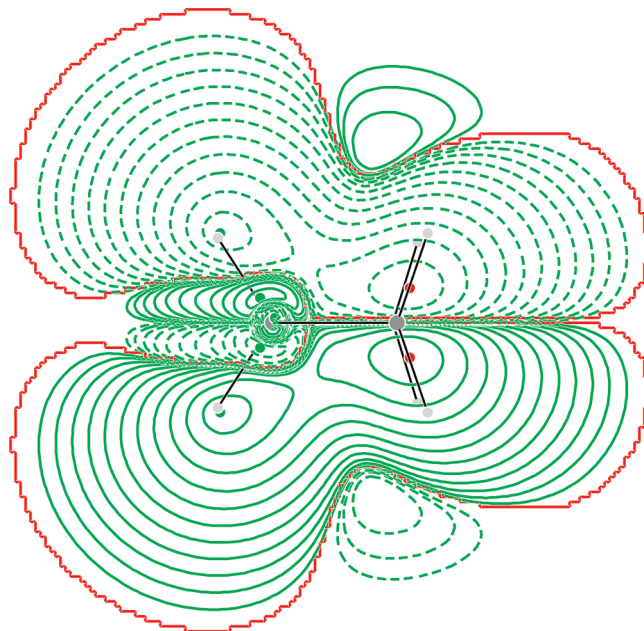


Figure 10. Contour plot of the cross-section of the modulus $|\mathbf{J}^{\mathbf{B}}|$ of the current density on the xz plane spanned by the $C_2 \equiv x$ and the $C_3 \equiv z$ symmetry axes, bisecting a C–C bond. The applied magnetic field B_x lies on the plot plane. The net integrated current strength on the lower (upper) domain with red boundary is $+15.663$ (-15.663) nA/T.

It is this current that could be referred to as a ring current in the conventional meaning.^{7,15} However, it should be emphasized that the shape of the total current density field is that of Figure 1 of the present study, and, accordingly, that both carbon and hydrogen nuclei lie inside a *torus of delocalized currents*. Plots of the shielding density Σ_{zz}^{H} , eq 7 on four planes, that of the C nuclei at $z = 0$, the H nuclei at $z = 1.702$ bohr, and two intermediate planes at 0.6 and 1.2 bohr are shown in Figure 11. To identify the contributions to proton shielding provided by the delocalized currents and by the local vortices flowing about the sp^3 orbitals forming the C–H bonds, for each plane in that figure, contour levels of Σ_{zz}^{H} are superimposed to the current density streamlines.

On the σ_h plane pairs of steep up and down spikes of the shielding density function, marked by green and red contours, respectively, are observed about carbon atoms. Their contributions to σ_{zz}^{H} are virtually vanishing due to cancellation of effects within each pair. On the same plane, a relatively large shielding region extends over the delocalized current domain, whose contribution can be considered minor, or negligible, due to its small magnitude. On going from the plane of the C to that of the H nuclei, both local and nonlocal current domains provide increasingly higher shielding effects of comparable size. A restricted area of deshielding is confined within the domain of local flow. On account of these results, the σ_{zz}^{H} component is determined by both local and nonlocal currents providing contributions of almost the same order of magnitude. The contribution of the delocalized flow seems slightly dominant owing to its wider extension.

The current density streamlines on the σ_h plane, and three parallel planes at distance 0.2, 0.5, and 0.8 bohr, are superimposed to the out-of-plane component of the shielding density functions, Σ_{zz}^{C} for carbon in Figure 12 and Σ_{zz}^{H} for a

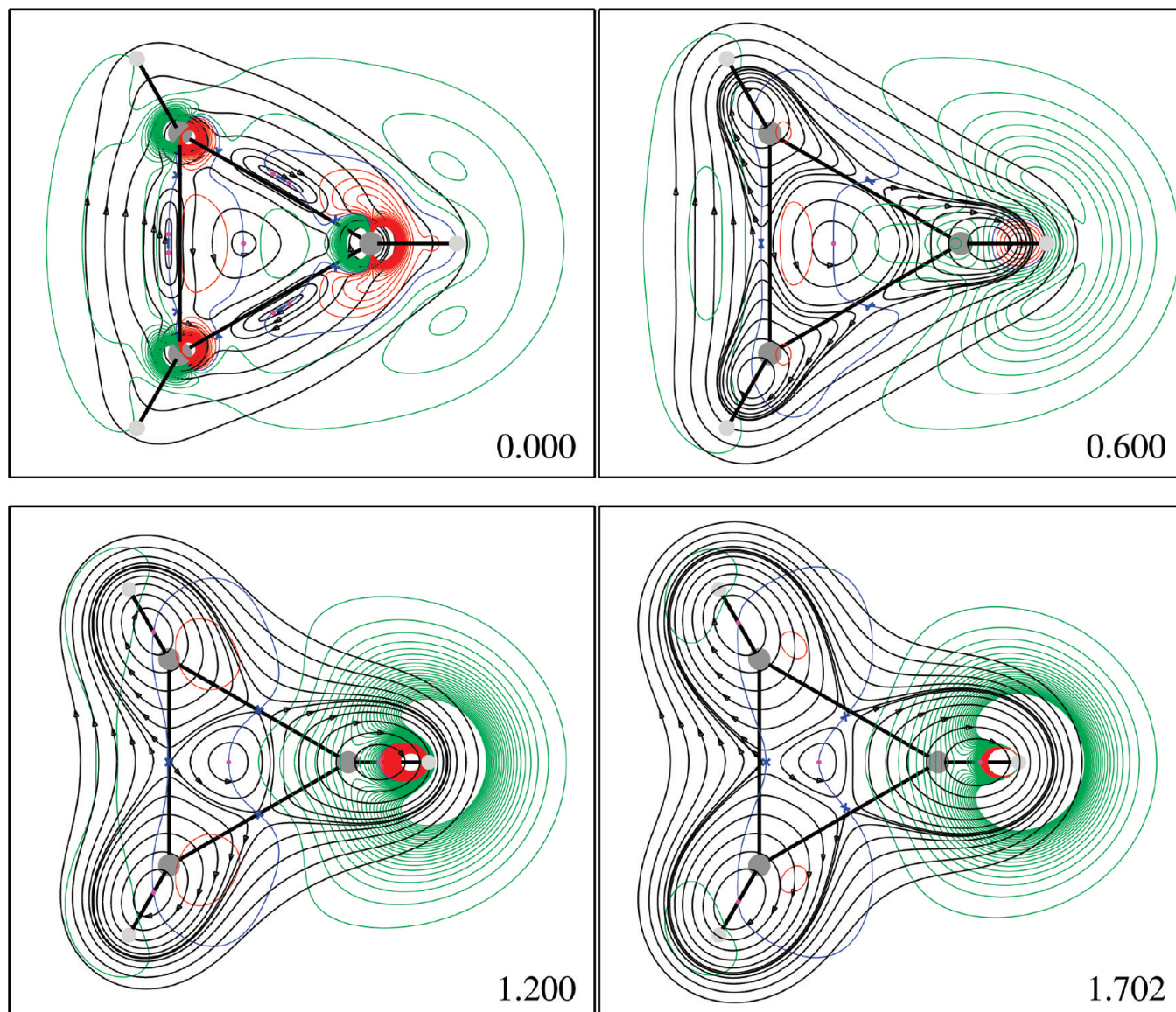


Figure 11. Streamlines of \mathbf{J}^B and corresponding contours of the magnetic shielding density Σ_{zz}^H in au, for an applied field of 1 au, on the σ_h plane and on three parallel planes at distance (in bohr) 0.6, 1.2, and 1.702 (the plane of the H nuclei). Green (red) contours denote shielding (deshielding). A blue nodal line connects points at which the angle between the local streamline and the direction to the probe is 0 or π , passing through the (2,0) vortex and saddle points on the plot plane, where $|\mathbf{J}^B|$ vanishes. The nodal line contains also the (3, ± 1) saddle nodes and foci on the σ_h plane. The shielding contributions, which arise from the σ -ring currents on the σ_h plane, are very small, those provided by the vortex flowing around the C–H bond and by the delocalized currents increase on the planes at 0.6 and 1.2 bohr and on the plane of the hydrogen nuclei. On these planes, the extension of the domains of delocalized flow providing shielding contributions is higher than that of the local vortex, implying that the former plays a slightly major role. Truncated min, max, and step = -0.05 , 0.05 , and 0.002 au.

ghost nucleus I on the center of mass in Figure 13. Let us first consider the Σ_{zz}^C shielding density plot of Figure 12. On the σ_h plane, a major contribution is provided by the diatropic vortex about the carbon atom, where a huge shielding spike is observed, reaching its maximum value within the domain of local current density flow. A sizable contribution arises also from the nearby domain of delocalized current. Pairs of nearly canceling up and down spikes are found on the other carbon atoms. A similar pattern is observed on the plane at 0.2 au. The contributions from the vortex flowing about the C–H bond and from the delocalized currents outside of it decrease on increasing the distance from σ_h . As shown in Figure 12, on the planes at 0.5 and 0.8 bohr, local and

nonlocal domains of flow seem to yield contributions of similar magnitude to σ_{zz}^C . Therefore, one can assert that the out-of-plane component of the carbon shielding is determined by a dominant local contribution and a smaller, non-negligible delocalized contribution.

Then let us consider the zz component of the central shielding density of Figure 13. On the σ_h plane, pairs of up and down spikes, corresponding to furthest and closest parts of the local C–H bond vortices, are observed about the carbon atoms. They yield negligibly small shielding contributions due to quasi-cancellation for each pair. The strong-deshielding zone confined inside the carbon ring is surrounded by a comparatively weaker-shielding region, spreading

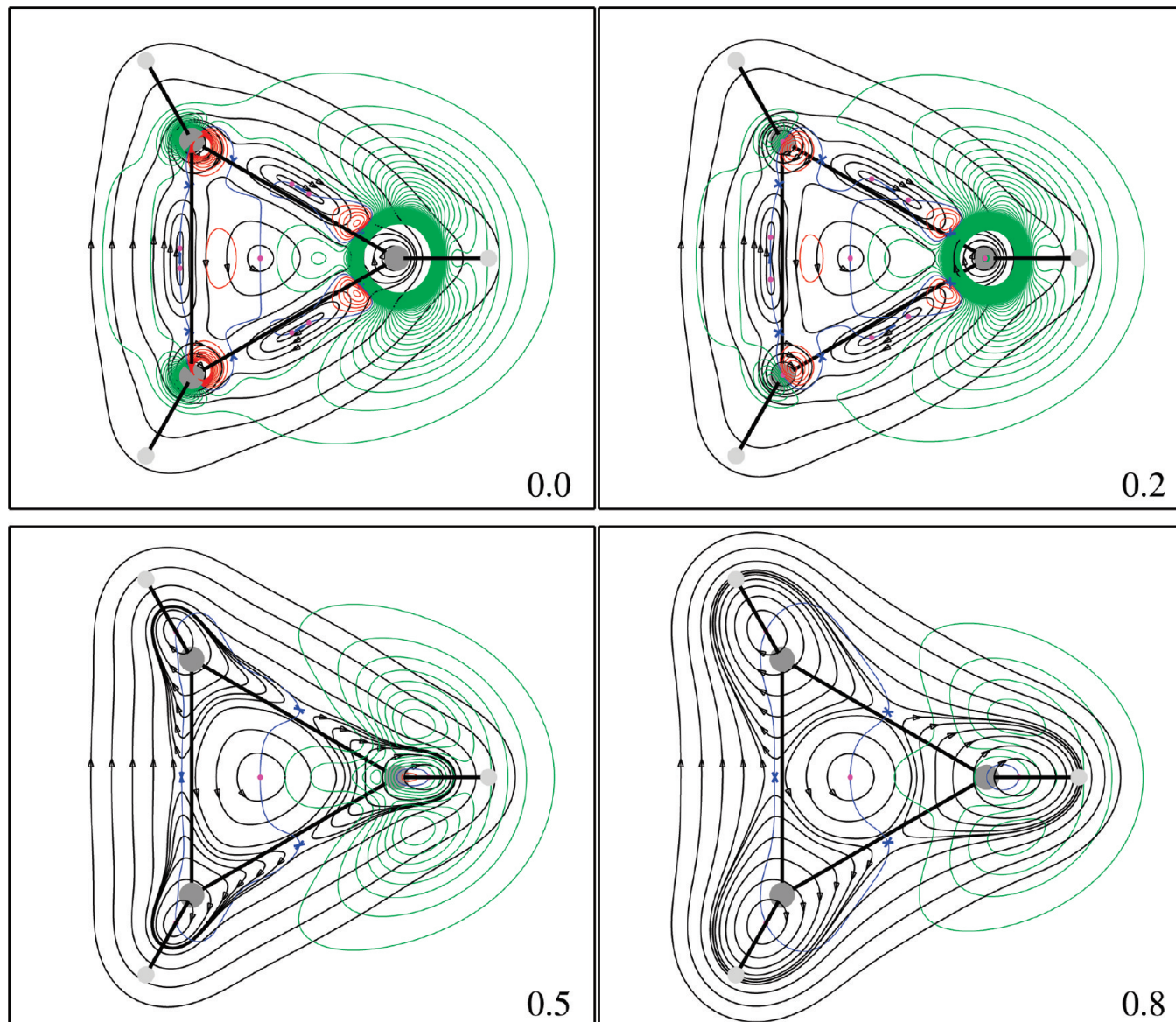


Figure 12. Streamlines and corresponding contours of the magnetic shielding density Σ_{zz}^C , in au, for an applied field of 1 au, on the plane of carbon nuclei and on three planes at distance (in bohr) 0.2, 0.5, and 0.8. The color code is the same as in Figure 11. The contributions arising from the vortex flowing around the C–H bond and from the delocalized currents outside of it decrease on increasing the distance from σ_h . Truncated min, max, and step = -0.10 , 1.00 , and 0.010 au.

however over a much larger area, up to the outer reaches of the σ_h plane, and reaching local maximum values just outside of the C–C bond directions, within the delocalized current domain.

A qualitatively similar pattern is observed on the plane at 0.2 bohr in Figure 13. The magnitude of Σ_{zz}^{CM} decays quite rapidly on more distant planes, relative maxima being still observed for the 0.5 and 0.8 planes in Figure 13, in the region of delocalized regime, in front of the C–C bonds. Therefore, it can be concluded that the large positive value of the out-of-plane component, $\sigma_{zz}^{CM} = 32.7$ ppm,²⁴ is determined to a large extent by the dominant shielding contribution provided by the delocalized current flow. Smaller but sizable deshielding contributions are given by the internal paratropic circulation.

8. Concluding Remarks

Spatial models of the current density field sustained by the electrons of the cyclopropane molecule in the presence of magnetic fields applied in the directions of either the C_3 or the C_2 symmetry axes, completing and partially revising that recently reported,²⁴ have been constructed. These models show that a magnetic field $B_{||}$ along C_3 induces localized vortices enclosed in a toroidal region of delocalized flow, which can be referred to as a spatial “ring current”, circulating beyond the skeleton of C and H nuclei and extending above and below the σ_h plane for more than 1.702 bohr, i.e., the plane of the hydrogen nuclei. Also a magnetic field B_{\perp} along C_2 induces intense delocalized flow outside of a set of localized vortices, with maximum intensity of

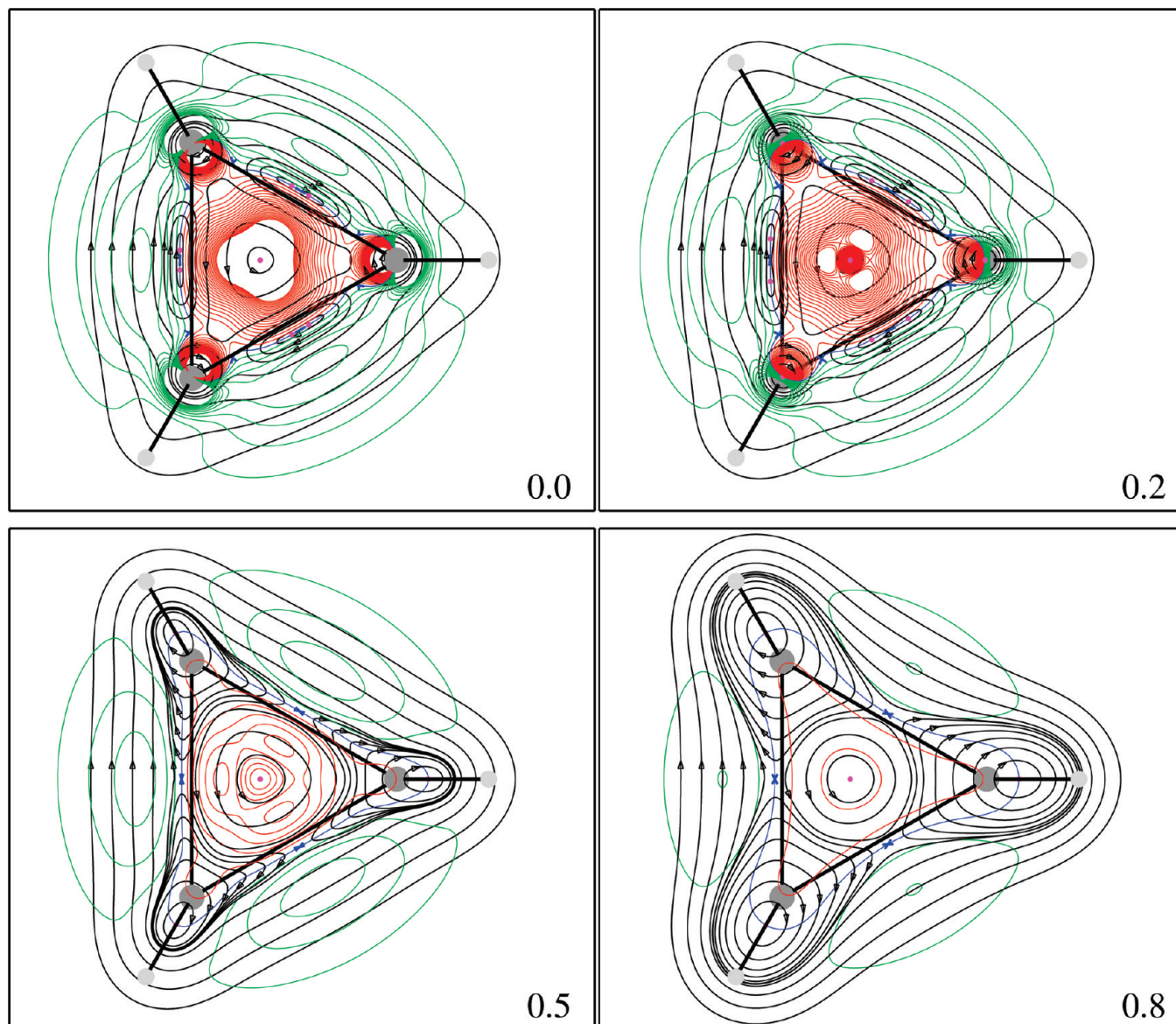


Figure 13. Streamlines and corresponding contours of the magnetic shielding density Σ'_{zz} , in au, for an applied field of 1 au, of a ghost nucleus I at the center of mass, on the plane of carbon nuclei and on three planes at distance (in bohr) 0.2, 0.5, and 0.8. The color code is the same as in Figure 11. The contributions to σ'_{zz} arising from the vortices flowing around the C–H bonds decrease on increasing the distance from σ_h . The intensity of delocalized currents beyond the methylene groups analogously decreases. On higher planes, small contributions to σ'_{zz} only arise from delocalized flow outside of CH₂–CH₂ moieties. Truncated min, max, and step = $-0.20, 0.10,$ and 0.010 au.

the same magnitude as that observed for $B_{||}$ applied in the C_3 direction.

Our ring current models have been compared with others previously reported,^{19,26} displaying current flow only in the plane of the carbon nuclei, and they have been tested by checking their ability to rationalize measurable quantities, such as the components of magnetic susceptibility and shielding of hydrogen and carbon nuclei. The magnetic shielding of a probe at the center of mass has also been interpreted. The main conclusions, confirming and widening those of a previous paper,²⁴ are:

- (1) In the presence of a magnetic field \mathbf{B} perpendicular to the plane of the carbon nuclei, the induced $\mathbf{J}^{\mathbf{B}}$ field contains four whirlpools extending for more than 10 bohr, above and below the σ_h plane, as shown in Figure 1. The central vortex, rotating about the C_3 axis,

is paratropic and of weaker intensity in comparison with the three strong diatropic vortices sustained by the sp^3 hybrid carbon orbitals forming the C–H bonds. This set of vortices is enclosed within a large domain of torus-shaped diatropic flow, delocalized around the whole skeleton of C and H nuclei. Comparatively higher intensities of this delocalized flow are observed along an annulus of “ring currents” originating in the C–C bent bonds, flowing outside of the carbon nuclei and splitting into two streams along the C–H bonds.

- (2) The $\mathbf{J}^{\mathbf{B}}$ field of C_nH_n cyclic planar systems sustaining delocalized diatropic π -currents contains n diatropic vortices originating at two points on C_n , equally spaced with respect to the center of mass (at $\approx \pm 2.5$ bohr in benzene)⁴⁸ and passing through the C–C bonds. At variance with this typical pattern, saddle regime is

observed about the midpoint of C–C bonds in cyclopropane. Spiralling trajectories connect a set of 18 foci and 6 saddle nodes.

- (3) The near Hartree–Fock average shielding $\sigma_{\text{av}}^{\text{H}} = 32.1$ is dominated by $\sigma_{\text{zz}}^{\text{H}} = 37.0$ ppm. The other calculated components are $\sigma_{\text{xx}}^{\text{H}} = 33.3$ and $\sigma_{\text{yy}}^{\text{H}} = 26.1$ ppm. The annular flow on the σ_h plane, involving cyclic σ -electron delocalization among the three carbon atoms, referred to as a ring current by Dale Poulter et al.,⁷ provides minor contributions to the high-field chemical shift $\delta^{\text{H}} = \sigma_{\text{av}}^{\text{H}}(\text{TMS}) - \sigma_{\text{av}}^{\text{H}}$ from tetramethylsilane (TMS). In fact, the proton shielding is determined to comparable extent by the delocalized currents and the local vortices about the CH bonds. As the major contribution of the delocalized current to $\sigma_{\text{zz}}^{\text{H}}$ comes from electron flow on planes close to the H nucleus, a simplified RCM should more properly consider a current loop on the plane of the hydrogen nuclei, rather than a circuit around the carbon nuclei on the σ_h plane.
- (4) The local diatropic vortices sustained by the sp^3 hybrid orbital cause also a major diamagnetic shift of the out-of-plane component $\sigma_{\text{zz}}^{\text{C}} = 236.0$ ppm of the carbon shielding, a corresponding increase of the average $\sigma_{\text{av}}^{\text{C}} = 198.1$ ppm and a sizable anisotropy $\Delta\sigma^{\text{C}} = \sigma_{\text{zz}}^{\text{C}} - (\sigma_{\text{xx}}^{\text{C}} + \sigma_{\text{yy}}^{\text{C}})/2 = -56.9$ ppm.
- (5) A magnetic field B_{\perp} parallel to a two-fold symmetry axis induces strong delocalized currents circulating about the entire molecule, sustained by the local charge distribution (the σ -electrons of the C–C “banana bond” and of the C–H bonds), as shown in Figure 6. The current susceptibility of the $T\sigma_v \equiv xz$ half-plane, for a magnetic field in the $C_2 \equiv x$ direction, is 15.7 nA/T, see Figure 10, that is, ≈ 1.5 times higher than that calculated for B_{\parallel} .
- (6) The noncanonical “ring current without a carbon ring” circulating about the C_2 axis enhances the $\sigma_{\perp}^{\text{CM}}$ in-plane component of the shielding tensor of a probe placed in the center of mass. The near Hartree–Fock value of $\sigma_{\perp}^{\text{CM}}$ is 50.9 ppm, that is ≈ 18 ppm bigger than out-of-plane $\sigma_{\parallel}^{\text{CM}} = 32.7$ (which would be the relevant quantity for an assessment of σ -diatropicity), thus providing a dominant contribution to the average $\sigma_{\text{av}}^{\text{CM}} = (\sigma_{\parallel}^{\text{CM}} + 2\sigma_{\perp}^{\text{CM}})/3 = 44.9$ ppm. Therefore, neither the average NICS, defined as $-\sigma_{\text{av}}^{\text{CM}}$,¹³ nor $\text{NICS}_{\parallel} = -\sigma_{\parallel}^{\text{CM}}$ ^{66,67} are reliable quantifiers of σ -diatropicity for cyclopropane. Furthermore, if $\sigma_{\parallel}^{\text{CM}}$ were preferred as an appropriate measure of magnetic aromaticity, one should admit that the cyclopropane molecule is even more diatropic when exposed to a magnetic field directed like a C_2 symmetry axis.
- (7) Whereas $\sigma_{\perp}^{\text{CM}} > \sigma_{\parallel}^{\text{CM}}$, $\chi_{\perp} < \chi_{\parallel}$ in cyclopropane. The peculiarity of these results, seemingly in contrast with one another, is understood by means of our RCM, recalling that the element of induced magnetic field at the position of a probe depends on the second inverse power of its distance from an element of current. The in-plane component of the shielding at the center of mass is mainly biased by the strong diatropic ring current nearest to it displayed in Figure

6. For instance, the plot of the isoshielding surface of Σ'_{xx} with value 0.05 in Figure 8, for a ghost atom at the midpoint of a C–C bond, shows shape and size of the shielding basin which provides big contributions to σ'_{\perp} . The out-of-plane component χ_{\parallel} of the magnetic susceptibility, a quantity depending on the intensity of the ring currents and on the area enclosed in the ring-current loop,⁶³ samples the whole molecular domain. As such, it can generally be considered as a more reliable measure of global diatropicity.

These results show that the basic motif of the well-established ad hoc model first proposed by Dewar¹⁵ to explain chemical shifts of proton magnetic resonance is to some extent confirmed and completed by that proposed in this work. The statement that cyclopropane is an archetypal σ -aromatic system only on the basis of the isotropic NICS value has been demonstrated to be unsatisfactory, as much of the NICS arises from the delocalized current flowing around the C_2 axes. Such a current is consistent with a significant nonlocal contribution to $\Delta\chi$, and it is useful to explain the surprising results reported by Benson and Flygare,^{5,23} cited in Section 1, as well as the conflicting conclusions on relative amounts of delocalization arrived at by considering either χ_{av} or $\Delta\chi$.

In fact, in Section 1, we evaluated the nonlocal contributions to χ_{av} and $\Delta\chi$, that is, $\chi_{\parallel}^{\text{nonloc}} = -14.5$ and $\chi_{\perp}^{\text{nonloc}} = +1.9$ ppm erg $\text{G}^{-2} \text{mol}^{-1}$, respectively. The estimated positive nonlocal contribution to χ_{\perp} depends on the assumption of sp^3 hybridization of carbon valence orbitals.

On the other hand, if one accepts the Walsh idea that cyclopropane might be portrayed as a π -complex of one ethylene and one methylene fragments,^{74,75} assuming sp^2 hybridization, using the atomic Pascal terms $\chi_{\text{C}} = -6.00$, $\chi_{\text{H}} = -2.93$, the correction 5.5 erg $\text{G}^{-2} \text{mol}^{-1}$ for a C=C bond reported by Bain and Berry,²¹ and the anisotropy $\Delta\chi = 4.4 \pm 0.4$ for an sp^2 carbon from Benson and Flygare,²² the nonlocal contribution to the average susceptibility becomes $\chi_{\text{av}}^{\text{nonloc}} = -39.2 - 3 \times (-6.00) - 6 \times (-2.93) - 5.5 = -9.1$, that to the anisotropy is $\Delta\chi^{\text{nonloc}} = -11.6 - 3 \times 4.4 = -24.8$, so that $\chi_{\parallel}^{\text{nonloc}} = -25.7$ and $\chi_{\perp}^{\text{nonloc}} = -0.9$ ppm erg $\text{G}^{-2} \text{mol}^{-1}$.

Therefore, the nonlocal contribution to χ_{\perp} becomes negative and smaller than the sum of the experimental errors. The fact that Pascal's constants are able to account for the effects of a delocalized current in cyclopropane, as well as in ethylene, is analogous to what happens in conjugated noncyclic hydrocarbons, where the stabilization energies can be obtained in terms of additive terms.⁷⁶

This finding might imply that a description in terms of sp^2 , instead of sp^3 , carbon hybrids^{74,75} would be preferable, which seems, to some extent, consistent with the model of the current density reported in Figures 1 and 6. In particular, the intense electron flow about a C_2 symmetry axis is fully compatible with the Walsh model of cyclopropane as a π -complex of one ethylene moiety and one methylene fragment.^{74,75}

Acknowledgment. R.C. thanks the Fund for Scientific Research (F.R.S.-FNRS) for his Research Fellow position. This work has been supported by the Belgian Government

(IUAP N P06-27 Functional Supramolecular Systems) and by a scientific collaboration between Wallonie-Bruxelles International, the FNRS, and the Italian Department of International Affairs. Financial support from the Fondazione Cassa di Risparmio di Modena and from the Italian MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca) via FARB funds of the University of Salerno is gratefully acknowledged.

Supporting Information Available: BK and FBL ring current model. Nuclear magnetic shielding of carbon and hydrogen from the CTOCD/6-31G** calculation. Magnetic susceptibilities from the CTOCD/6-31G** calculation. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Lacher, J. R.; Pollock, J. W.; Park, J. D. *J. Chem. Phys.* **1952**, *20*, 1047–1048.
- (2) Wiberg, K.; Nist, B. J. *J. Am. Chem. Soc.* **1961**, *83*, 1226–1230.
- (3) Patel, D. J.; Howden, M. H.; Roberts, J. D. *J. Am. Chem. Soc.* **1963**, *85*, 3218–3223.
- (4) Burke, J. J.; Lauterbur, P. C. *J. Am. Chem. Soc.* **1964**, *86*, 1870–1871.
- (5) Benson, R. C.; Flygare, W. H. *J. Chem. Phys.* **1969**, *51*, 3087–3096.
- (6) Bley, W.-R. *Mol. Phys.* **1971**, *20*, 491–501.
- (7) Dale Poulter, C.; Boikess, R. S.; Brauman, J. I.; Winstein, S. *J. Am. Chem. Soc.* **1972**, *94*, 2291–2296.
- (8) Hahn, R. C.; Howard, P. C. *J. Am. Chem. Soc.* **1972**, *94*, 3143–3148.
- (9) Aldrich, P. D.; Kukolich, S. G.; Campbell, E. J.; Read, W. G. *J. Am. Chem. Soc.* **1983**, *105*, 5569–5576.
- (10) Lukins, P. B.; Laver, D. R.; Buckingham, A. D.; Ritchie, G. L. D. *J. Phys. Chem.* **1985**, *89*, 1309–1312.
- (11) Jiao, H.; Nagelkerke, R.; Kurtz, H. A.; Williams, R. V.; Borden, W. T.; von Ragué Schleyer, P. *J. Am. Chem. Soc.* **1997**, *119*, 5921–5929.
- (12) Flygare, W. H. *Chem. Rev.* **1974**, *74*, 653–687.
- (13) von Ragué Schleyer, P.; Maerker, C.; Dransfeld, A.; Jiao, H.; van Eikema Hommes, N. J. R. *J. Am. Chem. Soc.* **1996**, *118*, 6317–6318.
- (14) Jensen, M. Ø.; Hansen, A. E. *Adv. Quantum Chem.* **1999**, *35*, 193–215.
- (15) Dewar, M. J. S. *J. Am. Chem. Soc.* **1984**, *106*, 669–682.
- (16) Tori, K.; Kitahonoki, K. *J. Am. Chem. Soc.* **1965**, *87*, 386–387.
- (17) Forsen, S.; Norin, T. *Tetrahedron Lett.* **1964**, *5*, 2845–2849.
- (18) Graham, J. D.; Rogers, M. T. *J. Am. Chem. Soc.* **1962**, *84*, 2249–2252.
- (19) Fowler, P. W.; Baker, J.; Lillington, M. *Theor. Chem. Acc.* **2007**, *118*, 123–127.
- (20) Wu, W.; Ma, B.; Wu, J. I.-C.; von Ragué Schleyer, P.; Mo, Y. *Chem.—Eur. J.* **2009**, *15*, 9730–9736.
- (21) Bain, G. A.; Berry, J. F. *J. Chem. Educ.* **2008**, *85*, 532–536.
- (22) Benson, R. C.; Flygare, W. H. *J. Chem. Phys.* **1973**, *58*, 2366–2372.
- (23) Benson, R. C.; Flygare, W. H. *J. Chem. Phys.* **1973**, *58*, 2651–2652.
- (24) Pelloni, S.; Lazzeretti, P.; Zanasi, R. *J. Phys. Chem. A* **2007**, *111*, 8163–8169.
- (25) Lazzeretti, P.; Zanasi, R. *J. Am. Chem. Soc.* **1983**, *105*, 12–15.
- (26) Bader, R. F. W.; Keith, T. A. *J. Chem. Phys.* **1993**, *99*, 3683–3693.
- (27) Fliegl, H.; Sundholm, D.; Taubert, S.; Jusélius, J.; Klopper, W. *J. Phys. Chem. A* **2009**, *113*, 8668–8676.
- (28) Exner, K.; von Ragué Schleyer, P. *J. Phys. Chem. A* **2001**, *105*, 3407–3416.
- (29) Moran, D.; Manoharan, M.; Heine, T.; von Ragué Schleyer, P. *Org. Lett.* **2003**, *5*, 23–26.
- (30) Dewar, M. J. S.; McKee, M. L. *Pure Appl. Chem.* **1980**, *52*, 1431–1441.
- (31) Minkin, V. I.; Glukhovtsev, M. N.; Simkin, B. Y. *J. Mol. Struct. (THEOCHEM)* **1988**, *181*, 93–110.
- (32) Li, Z.-H.; Moran, D.; Fan, K.-N.; von Ragué Schleyer, P. *J. Phys. Chem. A* **2005**, *109*, 3711–3716.
- (33) Monaco, G.; Zanasi, R.; Pelloni, S.; Lazzeretti, P., in preparation.
- (34) Bettinger, H. F.; Pak, C. H.; Xie, Y. M.; von Ragué Schleyer, P.; Schaefer, H. F. *J. Chem. Soc. Perkin Trans. 2* **1999**, 2377–2381.
- (35) Sauers, R. R. *Tetrahedron* **1998**, *54*, 337–348.
- (36) Reyn, J. W. Z. *Angew. Math. Physik* **1964**, *15*, 540–557.
- (37) Lazzeretti, P. Ring Currents. In *Progress in Nuclear Magnetic Resonance Spectroscopy*; Eds: Emsley, J. W., Feeney, J., Sutcliffe, L. H. Elsevier: 2000; Vol. 36, pp 1–88.
- (38) Gomes, J. A. N. F. *J. Chem. Phys.* **1983**, *78*, 4585–4591.
- (39) Gomes, J. A. N. F. *Phys. Rev. A* **1983**, *28*, 559–566.
- (40) Gomes, J. A. N. F. *J. Mol. Struct. (THEOCHEM)* **1983**, *93*, 111–127.
- (41) Keith, T. A.; Bader, R. F. W. *J. Chem. Phys.* **1993**, *99*, 3669–3682.
- (42) Collard, K.; Hall, G. G. *Int. J. Quantum Chem.* **1977**, *XII*, 623–637.
- (43) Bader, R. F. W. *Atoms in Molecules—A Quantum Theory*; Oxford University Press: Oxford, U.K., 1990.
- (44) Milnor, J. W. *Topology from the Differentiable Viewpoint*; University of Virginia Press: Charlottesville, VA, 1997.
- (45) Guillemin, V.; Pollack, A. *Differential Topology*; Prentice-Hall: Englewood Cliffs, NJ, 1974.
- (46) The topological index ι counts the number of times that the current density vector $\mathbf{J}^{\mathbf{B}}$ rotates completely, while one walks counterclockwise around a circle of radius ϵ so small that $\mathbf{J}^{\mathbf{B}}$ has no zeroes inside except the SP at its center. The topological index ι of a saddle (vortex) line is -1 ($+1$). Both SPs have $(r, s) = (2, 0)$.
- (47) Keith, T. A.; Bader, R. F. W. *Chem. Phys. Lett.* **1993**, *210*, 223–231.
- (48) Pelloni, S.; Faglioni, F.; Zanasi, R.; Lazzeretti, P. *Phys. Rev. A: At., Mol., Opt. Phys.* **2006**, *74*, 012506.

- (49) Pelloni, S.; Lazzeretti, P. *J. Chem. Phys.* **2008**, *128*, 194305-1–194305-10.
- (50) Keith, T. A.; Bader, R. F. W. *Can. J. Chem.* **1996**, *74*, 185–200.
- (51) Bader, R. F. W.; Keith, T. A. *Int. J. Quantum Chem.* **1996**, *60*, 373–379.
- (52) Pelloni, S.; Lazzeretti, P. *Theor. Chem. Acc.* **2007**, *117*, 903–913.
- (53) Pelloni, S.; Lazzeretti, P.; Zanasi, R. *J. Phys. Chem. A* **2007**, *111*, 3110–3123.
- (54) Gomes, J. A. N. F.; Mallion, R. B. *Chem. Rev.* **2001**, *101*, 1349–1383.
- (55) For detailed inspection, a graphic software is delivered, which can be used to rotate and to blow up this figure in three-dimensional space. The LINUX and WINDOWS versions of the graphic code used to obtain three-dimensional representations of the stagnation graph and current density vector field of a series of molecules can be downloaded at <https://theochem.chimfar.unimo.it/VEDO3/>.
- (56) Soncini, A.; Lazzeretti, P.; Zanasi, R. *Chem. Phys. Lett.* **2006**, *421*, 21.
- (57) Monaco, G.; Zanasi, R. *J. Chem. Phys.* **2009**, *131*, 044126.
- (58) Mitchell, R. H. *Chem. Rev.* **2001**, *101*, 1301–1315.
- (59) Hirschfelder, J. O. *J. Chem. Phys.* **1978**, *68*, 5151–5162.
- (60) Lazzeretti, P.; Malagoli, M.; Zanasi, R. *Chem. Phys. Lett.* **1994**, *220*, 299–304.
- (61) Jameson, C. J.; Buckingham, A. D. *J. Phys. Chem.* **1979**, *83*, 3366–3371.
- (62) Jameson, C. J.; Buckingham, A. D. *J. Chem. Phys.* **1980**, *73*, 5684–5692.
- (63) Since $\mathbf{J} = \rho\mathbf{v}$, $\mathbf{v} = d\mathbf{r}/dt$, $\rho d^3r = dq$, and $i = dq/dt$, the integrand in eq 2 is proportional to the current i and to the area element $\epsilon_{\alpha\beta\gamma}r_\gamma d\mathbf{r}_\beta$.
- (64) Lazzeretti, P.; Malagoli, M.; Zanasi, R. *Project Sistemi Informatici e Calcolo Parallelo*; Research Report 1/67; Consiglio Nazionale delle Ricerche (CNR): Rome, Italy, 1991.
- (65) Coriani, S.; Lazzeretti, P.; Malagoli, M.; Zanasi, R. *Theor. Chim. Acta* **1994**, *89*, 181–192.
- (66) Lazzeretti, P. *Phys. Chem. Chem. Phys.* **2004**, *6*, 217–223.
- (67) Corminboeuf, C.; Heine, T.; Seifert, G.; von Ragué Schleyer, P.; Weber, J. *Phys. Chem. Chem. Phys.* **2004**, *6*, 273–276.
- (68) Coulson, C. A.; Moffitt, W. E. *J. Chem. Phys.* **1947**, *15*, 151.
- (69) Pelloni, S.; Lazzeretti, P. *Chem. Phys.* **2009**, *356*, 153–163.
- (70) Monaco, G.; Zanasi, R. *AIP Conf. Proc.* **2009**, *1148*, 425–428.
- (71) Lin, Y.-C.; Sundholm, D.; Jusélius, J. *J. Chem. Theory Comput.* **2006**, *2*, 761–764.
- (72) Jusélius, J.; Sundholm, D.; Gauss, J. *J. Chem. Phys.* **2004**, *121*, 3952–3963.
- (73) Farnum, D. G.; Wilcox, C. F. *J. Am. Chem. Soc.* **1967**, *89*, 5379–5383.
- (74) Walsh, A. D. *Nature* **1947**, *159*, 165.
- (75) Walsh, A. D. *Nature* **1947**, *159*, 712–713.
- (76) Dewar, M. J. S.; Llano, C. D. *J. Am. Chem. Soc.* **1969**, *91*, 789–795.

CT100175J

Sparkle/PM6 Parameters for all Lanthanide Trications from La(III) to Lu(III)

Ricardo O. Freire[†] and Alfredo M. Simas^{*,‡}

Departamento de Química, Universidade Federal de Sergipe, 49.100-000, São Cristóvão, SE, Brazil and Departamento de Química Fundamental, Universidade Federal de Pernambuco, 50.740-540, Recife, PE, Brazil

Received April 11, 2010

Abstract: PM6 is the first semiempirical method to be released already parametrized for the elements of the periodic table, from hydrogen to bismuth ($Z = 83$), with the exception of the lanthanides from cerium ($Z = 58$) to ytterbium ($Z = 70$). In order to fill this gap, we present in this article a generalization of our Sparkle Model for the quantum chemical semiempirical calculation of lanthanide complexes to PM6. Accordingly, we present Sparkle/PM6 parameters for all lanthanide trications from La(III) to Lu(III). The validation procedure again considered only high-quality crystallographic structures and included 633 complexes. Sparkle/PM6 unsigned mean errors $UME_{(Ln-L)}s$, corresponding to all the interatomic distances between the lanthanide ion and the atoms directly coordinated to it, range from 0.066 to 0.086 Å for Gd(III) and Ce(III), respectively. These minimum and maximum $UME_{(Ln-L)}s$ across the lanthanide series are comparable to the Sparkle/AM1 values of 0.054 and 0.085 Å for Ho(III) and Pr(III), respectively, as well as to the values for Sparkle/PM3 of 0.064 and 0.093 Å for Gd(III) and Pr(III), respectively. Moreover, for all 15 lanthanide ions, these interatomic distance deviations follow a γ distribution within a 95% level of confidence, indicating that these errors appear to be random around a mean, freeing the model of systematic errors, at least within the validation set. Sparkle/PM6 presented here, therefore, broadens the range of applicability of PM6 to the coordination compounds of the rare earth metals.

Introduction

Parametric method number 6, PM6,¹ is the latest in a series of semiempirical methods which encompass MNDO,^{2,3} AM1,⁴ PM3,^{5–8} and RM1.⁹ The accuracy of PM6 in predicting enthalpies of formation, yielding an unsigned mean error of 4.4 kcal.mol⁻¹ for a representative set of 1373 compounds, exceeds those of Hartree–Fock (7.4 kcal.mol⁻¹) or B3LYP DFT (5.2 kcal.mol⁻¹) methods. PM6 has also been successfully used for modeling proteins and a variety of their properties.¹⁰ Moreover, PM6 has been further shown to be capable of reproducing the geometries and the enthalpies of formation of several solids with useful accuracy.¹¹

The Sparkle Model is a semiempirical approach to the quantum chemical calculation of lanthanide complexes, originally introduced by our research group in 1994.^{12,13} It replaces the lanthanide ions by a Coulombic charge of $+3e$, superimposed to a repulsive exponential potential of the form $\exp(-\alpha r)$, which was introduced to mimic the effect of the size of the ion. Thus, the Sparkle Model assumes that the angular effects of the f-orbitals are negligible and does not take them into account, being, thus, a spherically symmetric model. The Sparkle Model was improved in a subsequent article¹⁴ by the addition of two Gaussian functions to the core–core repulsion energy term, and by including the lanthanide mass, which allowed the calculation of vibrations and thermochemical quantities.¹⁵ Major and significant improvements to the parametrization procedure were then carried out, eventually leading to Sparkle/AM1,¹⁵ the first semiempirical quantum chemical model to be parametrized

* Corresponding author. E-mail: simas@ufpe.br.

[†] Universidade Federal de Sergipe.

[‡] Universidade Federal de Pernambuco.

for the whole lanthanide series.^{16–24} More recently, in order to allow the user a choice for the modeling of the organic motif of the complexes, Sparkle/PM3 was subsequently introduced.^{25–31}

The Sparkle Model was designed to reproduce the coordinating polyhedra of the complexes. That is because the geometry of a lanthanide coordination compound is its single most important feature for complex design.³² Indeed, for example, when designing a highly luminescent complex, one aims at achieving high-energy transfer rates from the ligands to the central metal ion, followed by a subsequent intense emission of light. For that purpose, the interaction between the ligands and the central metal ion can be described by the ligand field parameters, which can be calculated from the knowledge of the geometry of the coordination polyhedron. Within the simple overlap model,^{33,34} the ligand field parameters depend on the third, fifth, and seventh powers of the lanthanide–ligand interatomic distances, thus requiring an accurate knowledge of these distances. Likewise, for the design of contrast agents for magnetic resonance imaging, an accurate knowledge of the distance between the gadolinium ion and the oxygen atom of the coordinating water molecule is required. That is because the important dipolar relaxation mechanism has a dependency on the inverse sixth power of this distance. Again, any inaccuracies in this distance are greatly amplified when one tries to determine the relaxation rate of solvent protons, known as relaxivity.³⁵ A larger relaxivity implies that the required contrast agent may be administered in lower doses or that the imaging can be carried out in regions of lower contrast agent concentrations.

PM6 has been published and released already parametrized for 70 elements, from hydrogen to bismuth, with the exception of the lanthanides from cerium to ytterbium ($Z = 58$ and 70 , respectively).

Therefore, in order to broaden the range of applicability of PM6, we generalize in this article the Sparkle Model by introducing Sparkle/PM6 parameters for all lanthanide trications, from La(III) to Lu(III).

Results and Discussion

PM6 is a neglect of diatomic differential overlap (NDDO) method modified by the adoption of a slightly improved version of Voityuk's core–core diatomic parameters³⁶ to improve the predicted enthalpies of formation and geometries as well as rare gas interactions. On the other hand, the Sparkle Model does not require diatomic parameters, and therefore, for PM6, we maintained the same monatomic Sparkle core–core potential $E_N(A,B)$ with only two Gaussian functions, as fully described before.¹⁵

Although PM6, when released, already had parameters for lanthanum and lutetium, we decided to also make available PM6 sparkles for the trivalent ions of these atoms as well, to make the set consistent with Sparkle/AM1 and Sparkle/PM3.

Sparkle/PM6 parameters were obtained via the same parametrization procedure carried out to obtain the Sparkle/AM1¹⁵ and Sparkle/PM3 parameters.³¹ As such, we only used high-quality crystallographic structures

(R -factor $< 5\%$) obtained from the Cambridge Structural Database (CSD),^{37,38} and for the case of promethium, structures obtained by ab initio calculations,²¹ having selected a total of 633 complexes. From these, we took, as training sets, the same 15 subsets of 15 complexes for each lanthanide trication, previously chosen for the parametrization of Sparkle/AM1,^{16–24} and carried out the optimization following the same methodology as described before.¹⁵

The Sparkle/PM6 parameters found for all 15 lanthanide trications are shown in Table 1. In order to proceed with the validation, we used as geometry accuracy measures the average unsigned mean error for each complex i , UME_i , defined by

$$UME_i = \frac{1}{n_i} \sum_{j=1}^{n_i} |R_{ij}^{CSD} - R_{ij}^{calc}| \quad (1)$$

where n_i is the number of atoms in the coordination polyhedron of the lanthanide ion. Two cases were examined: (i) $UME_{(Ln-L)}$, where we considered only the interatomic distances between the lanthanide ion, Ln, and the atoms directly coordinated to it, L, and (ii) UMEs of all the edges of all faces of the pyramids defined by the lanthanide ion in the apex as well as all interatomic distances between all atoms of the coordination polyhedron.

Once again, it is important to assess whether Sparkle/PM6 provides a good and reliable representation of lanthanide complexes, free of systematic errors, at least within the validation set of complexes. For that to be true, the $UME_{(Ln-L)}$ s and the UMEs of all distances of all complexes of the test set should be randomly distributed about the mean, whose value can be used as a measure of the accuracy of the model. Therefore, $UME_{(Ln-L)}$ s and the UMEs should follow the probability density function of the γ distribution, since UMEs are positive and defined in the domain $(0, \infty)$. We then proceeded by obtaining a fit of the UMEs to a γ distribution function, from which the mean and variance could be determined. The quality of the γ distribution fit can be assessed via the one-sample nonparametric Kolmogorov–Smirnov test³⁹ in order to verify statistically whether the distribution of the UME values could be represented by a γ distribution indexed by the estimated mean and variance. If the p -value of the test is larger than 0.05, the γ distribution fit is indeed justified within a 95% confidence interval, and the mean and variance can be used as accuracy measures of the model.

The Supporting Information contains tables with the individual $UME_{(Ln-L)}$ s and UMEs for all 633 complexes. Table 2 summarizes these data for the $UME_{(Ln-L)}$ s for each lanthanide trication. The p -values of the γ distribution fits are all well above 0.05, implying they are all statistically valid. The average value of all $UME_{(Ln-L)}$ means is 0.0741 Å, whereas the average value of the respective variances is 0.0012 Å². One can see that these values are relatively constant throughout the table, with the maximum $UME_{(Ln-L)}$ being 0.0856 Å for Ce(III) and the minimum being 0.0663 Å for Gd(III). The corresponding low value for Pm(III), 0.0619 Å, is not strictly comparable with the others because

Table 1. Parameters for the Sparkle/PM6 Model for All Lanthanide Trications from La(III) to Lu(III)

Sparkle/PM6					
	La ³⁺	Ce ³⁺	Pr ³⁺	Nd ³⁺	Pm ³⁺
α	2.0955474333	2.1249588196	2.4693712260	4.1738480733	3.0374070006
GSS	55.7614959637	58.8260906171	58.3343604075	57.6974644976	59.5665725491
a_1	0.9198962192	1.7329167054	2.8321015232	1.1507966088	1.8134017776
b_1	7.1956586116	7.4140930052	7.1195524904	6.4949658378	9.0994056545
c_1	1.8688421745	1.7149281546	1.6208236553	1.5653255583	1.6148716177
a_2	0.3395617280	0.0764294472	0.0541169724	0.1889516026	0.2759193756
b_2	8.5194840290	8.4974750829	7.8230014741	10.9231117908	7.2120871121
c_2	3.1236739454	3.0778367381	3.1133411960	3.0169407149	3.0287226366
	Sm ³⁺	Eu ³⁺	Gd ³⁺	Tb ³⁺	Dy ³⁺
α	4.0858458124	2.0467722838	2.1346333468	2.5139941133	2.5510632015
GSS	56.8573294165	55.6632255486	56.8944696903	55.2205687662	55.8786332882
a_1	1.5645679440	0.2712333739	0.2517865588	0.5222813525	1.1809130487
b_1	6.4255324886	7.3743656586	8.7505991931	7.9527873623	8.9849822704
c_1	1.4885991013	1.7955662564	1.7313405711	1.7550018623	1.6756952638
a_2	0.1021969444	0.3493713916	0.1221903028	0.3099626210	0.4066395540
b_2	9.4102061689	7.7881047906	7.4979582981	6.6812787003	8.9799453811
c_2	3.1094973204	2.9632616015	2.9344373061	2.9759649920	2.9787279400
	Ho ³⁺	Er ³⁺	Tm ³⁺	Yb ³⁺	Lu ³⁺
α	3.4814819284	3.6603230421	2.3042905227	4.2104920412	3.2076166779
GSS	56.0010800433	58.4870426290	56.3699484190	56.3592753390	56.2871833842
a_1	0.3389541104	0.4687052850	0.7757838661	1.0542080628	0.6496316332
b_1	8.1824420999	9.3819581436	8.3570694122	8.5454710978	9.2468459960
c_1	1.6446707189	1.7306657473	1.6489766048	1.4993488570	1.5344631779
a_2	0.1333201849	0.2107436963	0.2905574744	0.1983232376	0.2355401411
b_2	8.7112042124	8.4256041357	7.6933919381	8.4702758246	7.3208383861
c_2	2.9809112221	2.7714227710	2.9316355211	2.8575372636	2.9270906286

Table 2. Means and Variances of the γ Distribution Fits for the UME_(Ln-L)s Computed for the N Complexes for Each Lanthanide Trication^a

lanthanide ion	N	UME _(Ln-L)		p -value
		mean (Å)	variance (Å ²)	
La ³⁺	73	0.0739	0.0011	0.732
Ce ³⁺	36	0.0856	0.0020	0.687
Pr ³⁺	47	0.0779	0.0014	0.994
Nd ³⁺	57	0.0744	0.0011	0.679
Pm ³⁺	15	0.0619	0.0015	0.947
Sm ³⁺	37	0.0748	0.0010	0.956
Eu ³⁺	88	0.0775	0.0013	0.159
Gd ³⁺	64	0.0663	0.0008	0.249
Tb ³⁺	35	0.0743	0.0007	0.961
Dy ³⁺	26	0.0798	0.0005	0.443
Ho ³⁺	28	0.0695	0.0006	0.838
Er ³⁺	38	0.0670	0.0028	0.761
Tm ³⁺	15	0.0734	0.0007	0.973
Yb ³⁺	44	0.0777	0.0013	0.993
Lu ³⁺	30	0.0778	0.0015	0.266

^aThe last column shows the p -values of the one-sample nonparametric Kolmogorov–Smirnov tests,³⁹ carried out for each lanthanide ion, in order to verify statistically that its value is above 0.05, indicating that the distribution of the UME_(Ln-L) values can indeed be represented by a γ distribution indexed by the estimated mean and variance. N refers to the number of complexes used in the comparison.

it was computed based on geometries obtained from ab initio calculations²¹ and not from crystallographic measurements, since Pm is a synthetic element which does not possess a stable nucleus.

Table 3 also summarizes the data for the UMEs for each lanthanide trication. In this case, the average value of all means is 0.192 Å, and the average value of all variances is 0.0066 Å². Once again, one can see that these values are relatively constant throughout the table, with the maximum

Table 3. Means and Variances of the γ Distribution Fits for the UMEs Computed for the N Complexes for Each Lanthanide Trication^a

lanthanide ion	N	UME		
		mean (Å)	variance (Å ²)	p -value
La ³⁺	73	0.213	0.0091	0.735
Ce ³⁺	36	0.190	0.0073	0.988
Pr ³⁺	47	0.212	0.0089	0.632
Nd ³⁺	57	0.198	0.0086	0.895
Pm ³⁺	15	0.165	0.0047	0.583
Sm ³⁺	37	0.188	0.0066	0.262
Eu ³⁺	88	0.195	0.0046	0.957
Gd ³⁺	64	0.186	0.0066	0.460
Tb ³⁺	35	0.191	0.0062	0.620
Dy ³⁺	26	0.204	0.0039	0.418
Ho ³⁺	28	0.214	0.0053	0.957
Er ³⁺	38	0.204	0.0102	0.948
Tm ³⁺	15	0.174	0.0048	0.892
Yb ³⁺	44	0.169	0.0041	0.666
Lu ³⁺	30	0.183	0.0084	0.819

^aThe last column shows the p -values of the one-sample nonparametric Kolmogorov–Smirnov tests,³⁹ carried out for each lanthanide ion, in order to verify statistically that its value is above 0.05, indicating that the distribution of the UME values can indeed be represented by a γ distribution indexed by the estimated mean and variance. N refers to the number of complexes used in the comparison.

UME being 0.214 Å for Ho(III) and the minimum being 0.169 Å for Yb(III). These data imply that Sparkle/PM6 is well balanced across the lanthanide series.

Figure 1 presents a histogram comparing UME_(Ln-L)s for all three Sparkle Models: Sparkle/AM1, Sparkle/PM3, and Sparkle/PM6. The UME_(Ln-L)s represent the accuracy of the model in terms of the distances between the lanthanide ion and the coordinated atoms, important for the design of both luminescent complexes and contrast agents for MRI. The

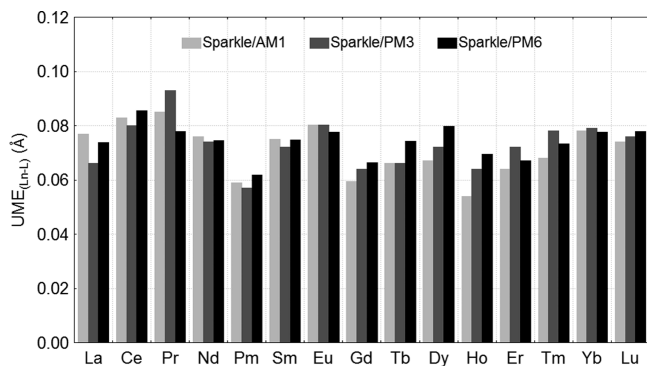


Figure 1. $UME_{(Ln-L)}$ s obtained using all three versions of the Sparkle Model: Sparkle/AM1, Sparkle/PM3, and Sparkle/PM6, for all complexes of the validation set and lanthanide trications, from La(III) to Lu(III). The UMEs are calculated as the average of the absolute value of the difference between the experimental and calculated interatomic distances between the lanthanide ion and the directly coordinating ligand atoms, summed for all complexes for each of the lanthanides.

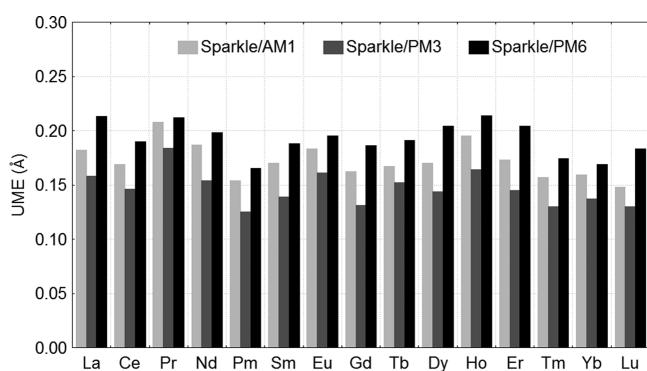


Figure 2. UMEs obtained using all three versions of the Sparkle Model: Sparkle/AM1, Sparkle/PM3, and Sparkle/PM6, for all complexes of the validation set and lanthanide trications, from La(III) to Lu(III). The UMEs are calculated as the average of the absolute value of the difference between the experimental and calculated interatomic distances between all atoms in the coordination polyhedron (lanthanide ion included), summed for all complexes for each of the lanthanides.

accuracy trends in this measure are similar for all three models across the lanthanide series, making them equivalent parametrizations. Figure 2 also presents a histogram comparing UMEs for all three Sparkle Models. In this case, there are more variations among the models, although their behavior is constant across the lanthanide series; the most accurate one being Sparkle/PM3, followed by Sparkle/AM1, and finally Sparkle/PM6. These trends probably reflect aspects of the original parametrizations of each method, not for the lanthanide but mainly for the types of atoms normally found in the coordination polyhedron.

In order to employ any of the Sparkle Models in MOPAC2009,⁴⁰ one must use the keyword SPARKLE together with the keyword of the chosen method: AM1, PM3, or PM6.⁴¹ Actually, to use Sparkle/PM6 the keyword SPARKLE only is normally sufficient, since PM6 is the default method of MOPAC2009. On the other hand, the PM6 article reports parameters for lanthanum and lutetium as regular PM6 atoms with orbitals, parameters which are also

Table 4. Sparkle/AM1, Sparkle/PM3, Sparkle/PM6, and PM6 Unsigned Mean Errors^a

Model	unsigned mean errors for specific types of distances (Å)					
	Ln-Ln	Ln-O	Ln-N	L-L'	Ln-L and Ln-Ln	Ln-L, Ln-Ln, and L-L'
Lanthanum(III)						
Sparkle/AM1	0.176	0.086	0.048	0.208	0.077	0.182
Sparkle/PM3	0.104	0.060	0.083	0.179	0.066	0.158
Sparkle/PM6	0.208	0.076	0.062	0.240	0.074	0.213
PM6	2.392	0.711	0.494	0.796	0.544	0.714
Lutetium(III)						
Sparkle/AM1	0.222	0.084	0.047	0.170	0.074	0.148
Sparkle/PM3	0.176	0.083	0.054	0.145	0.076	0.130
Sparkle/PM6	0.201	0.089	0.048	0.212	0.078	0.183
PM6	0.788	0.163	0.059	0.272	0.124	0.227

^a For all distances involving the central lanthanide ion, Ln, and the ligand atoms of the coordination polyhedron, L and L', and the specific cases when L is either oxygen or nitrogen for the 73 La(III) complexes and 30 Lu(III) complexes considered.

present and implemented in MOPAC2009. So, for the cases of lanthanum and lutetium, complexes can also be computed from pure PM6 parameters. Table 4 presents a comparison among all three Sparkle Models and the pure PM6 for such complexes with respect to several geometry accuracy measures. Clearly the geometry errors of pure PM6 for these two lanthanide atoms are a few times larger than the corresponding ones for all three Sparkle Models. Thus, only if properties other than geometries are required, the usage of pure PM6 for these two elements would be justified. However, due to the magnitude of the errors, even in these cases one could perhaps consider the possibility of optimizing the geometry with Sparkle/PM6 and then computing the other properties of interest with pure PM6 at the sparkle geometries.

As indicated above, Sparkle/PM6 is already implemented in MOPAC2009⁴⁰ and has been tested independently by Seitz and Alzakhem⁴² with respect to its ability to predict the average bond lengths of Ln-OH₂ for the technologically important central lanthanides, Ln = Eu, Gd, and Tb. These authors studied two classes of complexes: the first composed of pyridine-like ligands with 172 complexes, and the second featured ligands with the cyclen motif with 51 complexes. They concluded that Sparkle/AM1 is best for complexes with pyridine-like ligands, whereas Sparkle/PM6 outperforms the other two Sparkle Models in cyclen-derived species. This assertion clearly illustrates and justifies the importance of having all three Sparkle Models available because the individual characteristics of each underlying semiempirical method (AM1, PM3, or PM6) may be more applicable to one or another specific situation.

Conclusions

Sparkle/PM6 stands as another option in the suite of semiempirical models applicable to the quantum chemical calculation of lanthanide complexes. Sparkle/PM6 is an accurate and statistically valid tool for the prediction of the geometrical features of lanthanide coordination polyhedra and, by design, is expected to perform best with ligands with nitrogen or oxygen as coordinating atoms present in the vast majority of all coordination compounds of the trivalent rare earth metals.

Acknowledgment. We appreciate the financial support from the following Brazilian agencies, institutes, and networks: CNPq, FACEPE (Pronex), FAPITEC-SE, INAMI, and RENAMI. We are also grateful to Prof. A.E.A.Paixão for the use of the software Statistica. Finally, we gratefully acknowledge the Cambridge Crystallographic Data Centre for the Cambridge Structural Database.

Supporting Information Available: Instructions on how to run lanthanide complexes Sparkle calculations (<http://www.sparkle.pro.br>) with MOPAC 2009 (<http://openmopac.net>). Additional tables containing UME_(Ln-L)s and UMEs for all 633 complexes of the validation set. Additional histograms comparing all three Sparkle Models with respect to various classes of coordinating bonds. Sample MOPAC 2009 input (.mop) and output (.arc) files for one representative complex for each of the lanthanide ions, from La(III) to Lu(III). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173–1213.
- Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.
- Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4907–4917.
- Dewar, M. J. S.; Zuehlke, E. G.; Healy, E. F.; e Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209–220.
- Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221–264.
- Stewart, J. J. P. *J. Comput. Chem.* **1991**, *12*, 320–341.
- Anders, E.; Koch, R.; Freunshcht, P. *J. Comput. Chem.* **1993**, *14*, 1301–1312.
- Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101–1111.
- Stewart, J. J. P. *J. Mol. Model.* **2009**, *15*, 765–805.
- Stewart, J. J. P. *J. Mol. Model.* **2008**, *14*, 499–535.
- de Andrade, A. V. M.; da Costa, N. B., Jr.; Simas, A. M.; de Sá, G. F. *Chem. Phys. Lett.* **1994**, *227*, 349–353.
- de Andrade, A. V. M.; da Costa, N. B., Jr.; Simas, A. M.; de Sá, G. F. *J. Alloys Compd.* **1995**, *225*, 55–59.
- Rocha, G. B.; Freire, R. O.; da Costa, N. B., Jr.; de Sá, G. F.; Simas, A. M. *Inorg. Chem.* **2004**, *43*, 2346–2354.
- Freire, R. O.; Rocha, G. B.; Simas, A. M. *Inorg. Chem.* **2005**, *44*, 3299–3310.
- Freire, R. O.; da Costa, N. B., Jr.; Rocha, G. B.; Simas, A. M. *J. Organomet. Chem.* **2005**, *690*, 4099–4102.
- Freire, R. O.; Rocha, G. B.; Simas, A. M. *Chem. Phys. Lett.* **2005**, *411*, 61–65.
- Freire, R. O.; Rocha, G. B.; Simas, A. M. *J. Comput. Chem.* **2005**, *26*, 1524–1528.
- da Costa, N. B., Jr.; Freire, R. O.; Rocha, G. B.; Simas, A. M. *Polyhedron* **2005**, *24*, 3046–3051.
- da Costa, N. B., Jr.; Freire, R. O.; Rocha, G. B.; Simas, A. M. *Inorg. Chem. Commun.* **2005**, *8*, 831–835.
- Freire, R. O.; da Costa, N. B., Jr.; Rocha, G. B.; Simas, A. M., J. *J. Chem. Theory Comput.* **2006**, *2*, 64–74.
- Bastos, C. C.; Freire, R. O.; Rocha, G. B.; Simas, A. M. *J. Photochem. Photobiol., A* **2006**, *117*, 225–237.
- Freire, R. O.; Monte, E. V.; Rocha, G. B.; Simas, A. M. *J. Organomet. Chem.* **2006**, *691*, 2584–2588.
- Freire, R. O.; da Costa, N. B., Jr.; Rocha, G. B.; Simas, A. M. *J. Phys. Chem. A* **2006**, *110*, 5897–5900.
- Freire, R. O.; Rocha, G. B.; Simas, A. M. *Chem. Phys. Lett.* **2006**, *425*, 138–141.
- Freire, R. O.; Rocha, G. B.; Simas, A. M. *Chem. Phys. Lett.* **2007**, *441*, 354–357.
- da Costa, N. B., Jr.; Freire, R. O.; Rocha, G. B.; Simas, A. M. *J. Phys. Chem. A* **2007**, *111*, 5015–5018.
- Simas, A. M.; Freire, R. O.; Rocha, G. B. *Lect. Notes Comput. Sci.* **2007**, *4488*, 312–318.
- Freire, R. O.; da Costa, N. B., Jr.; Rocha, G. B.; Simas, A. M., J. *J. Chem. Theory Comput.* **2007**, *3*, 1588–1596.
- Simas, A. M.; Freire, R. O.; Rocha, G. B. *J. Organomet. Chem.* **2008**, *693*, 1952–1956.
- Freire, R. O.; Rocha, G. B.; Simas, A. M. *J. Braz. Chem. Soc.* **2009**, *20*, 1638–1645.
- Rodrigues, M. O.; da Costa, N. B., Jr.; de Simone, C. A.; Araujo, A. A. S.; Brito-Silva, A. M.; Paz, F. A. A.; de Mesquita, M. E.; Junior, S. A.; Freire, R. O. *J. Phys. Chem. B* **2008**, *112*, 4204–4212.
- Malta, O. L. *Chem. Phys. Lett.* **1982**, *87*, 27–29.
- Malta, O. L. *Chem. Phys. Lett.* **1982**, *88*, 353–356.
- Caravan, P.; Astashkin, A. V.; Raitsimring, A. M. *Inorg. Chem.* **2003**, *42*, 3972–3974.
- Voityuk, A. A.; Rösch, N. *J. Phys. Chem. A* **2000**, *104*, 4089–4094.
- Allen, F. H. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.
- Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 389–397.
- William, J. C. *Practical nonparametric statistics*; John Wiley & Sons: New York, 1971.
- Stewart; J. J. P. *MOPAC2009*, version 10.060W; Stewart Computational Chemistry: Colorado Springs, CO, 2009.
- MOPAC2009 Home Page; <http://openmopac.net/manual/index.html>. Accessed April, 2010.
- Seitz, M.; Alzakhem, N. *J. Chem. Inf. Model.* **2010**, *50*, 217–220.

Generalized Møller–Plesset Partitioning in Multiconfiguration Perturbation Theory

Masato Kobayashi,^{†,‡,§} Ágnes Szabados,[†] Hiromi Nakai,^{‡,||} and Péter R. Surján^{*,†}

Laboratory of Theoretical Chemistry, Institute of Chemistry, Eötvös University, H1518 Budapest POB 32, Hungary, Department of Chemistry and Biochemistry, School of Advanced Science and Engineering, Waseda University, Tokyo 169-8555, Japan, Department of Theoretical and Computational Molecular Science, Institute for Molecular Science, Okazaki 444-8585, Japan, and Research Institute for Science and Engineering (RISE), Waseda University, Tokyo 169-8555, Japan

Received April 12, 2010

Abstract: Two perturbation (PT) theories are developed starting from a multiconfiguration (MC) zero-order function. To span the configuration space, the theories employ biorthogonal vector sets introduced in the MCPT framework. At odds with previous formulations, the present construction operates with the full Fockian corresponding to a principal determinant, giving rise to a nondiagonal matrix of the zero-order resolvent. The theories provide a simple, generalized Møller–Plesset (MP) second-order correction to improve any reference function, corresponding either to a complete or incomplete model space. Computational demand of the procedure is determined by the iterative inversion of the Fockian, similarly to the single reference MP theory calculated in a localized basis. Relation of the theory to existing multireference (MR) PT formalisms is discussed. The performance of the present theories is assessed by adopting the antisymmetric product of strongly orthogonal geminal (APSG) wave functions as the reference function.

1. Introduction

Single-reference quantum chemical methods have achieved great success in describing molecular electronic structures at around equilibrium geometry. However, these methods fail in calculating systems which have near degeneracy around frontier orbitals, a situation often encountered at geometries far from equilibrium structures. For treating these latter systems, multireference (MR) variational theories have been proposed, such as multiconfigurational self-consistent field (MCSCF) theory,¹ complete active space self-consistent field (CASSCF) theory,² geminal-based theories including generalized valence bond (GVB) theory,³ or the antisymmetric product of strongly orthogonal geminals (APSG) theory.^{4–7} Although these methods can improve the descrip-

tion of degenerate systems qualitatively, they usually provide an insufficient amount of dynamic correlation energy, unless the variational space is extended to cover such a large portion of the configuration space, which in turn reduces the practical applicability of the approach. To achieve a significant inclusion of dynamic and static correlation at the same time it is well established to apply perturbation (PT) or coupled-cluster (CC) theories based on a multideterminantal wave function.

Multireference extension of PT theories has spawned a number of alternative formulations, the developments continuously being carried out. As a guiding rule, MR PT approaches can be categorized as being either (i) effective Hamiltonian theories with a model space of dimension higher than one (“perturb then diagonalize”)^{8,9} or (ii) theories that apply to a one-dimensional model space (“diagonalize then perturb”). Focusing on category ii, there is still a large variety of different formulations. For its obvious success in the realm of single-determinant dominated systems, the Møller–Plesset (MP) partitioning of standard Rayleigh–Schrödinger PT (the Fock operator playing the role of the unperturbed Hamilto-

* To whom correspondence should be addressed: E-mail: surjan@chem.elte.hu.

[†] Eötvös University.

[‡] Waseda University, Department of Chemistry and Biochemistry.

[§] Institute for Molecular Science.

^{||} Waseda University, RISE.

nian) was generalized to the MR case in particularly diverse ways. A common origin of several of these theories is the general expression of their zero-order Hamiltonian in the form

$$\hat{H}^{(0)} = E^{(0)}\hat{O} + \hat{P}\hat{F}\hat{P} \quad (1)$$

where \hat{O} is the projector corresponding to the one-dimensional space spanned by the reference function and $\hat{P} = 1 - \hat{O}$ is the projector complementary and orthogonal to \hat{O} . Specific theories differ in the definition of the Fockian \hat{F} , the form of projector \hat{O} , the definition of the reference energy $E^{(0)}$, the functions applied to span space \hat{P} , and the treatment of their incidental overlap. It is also common to apply a decoupled form of eq 1, as will be discussed below.

In the present study, we devise a novel PT scheme that operates with the general form of eq 1 of the zero-order Hamiltonian and can be considered as the extension of the MP partitioning to the previously introduced multiconfiguration PT (MCPT) framework.^{10,11} Previous variants of MCPT employed a diagonal zero-order Hamiltonian with a choice of zero-order energies. In the present formulation, this flexibility is left off by projecting the full Fockian into space \hat{P} according to eq 1. The zero-order Hamiltonian is non-Hermitian, due to the application of biorthogonal vector sets specific to MCPT. Two alternatives of handling the overlap between excited determinants and the reference function lead to two MCPT variants with the MP partitioning. One will be referred to as projected or pMCPT; the other will be called unprojected or uMCPT.

To avoid any confusion, we note that acronym “u” in uMCPT is not the shorthand commonly used for unrestricted orbitals. In the present work, we consider restricted orbitals throughout. In principle, the determinant-based formalism presented below makes the extension of the theory straightforward for unrestricted orbitals. Such an extension shows relations with the USSG (unrestricted strongly orthogonal singlet-type geminals)-based perturbation theory developed by Rassolov and co-workers¹² and may be achieved without violation of the spin-symmetry.¹³

In this report, we first present the extension of MP partitioning to the MCPT framework in section 2. This is followed by a separate short section, section 3, devoted to the question of size-consistency, followed by a survey of related formulations in section 4. Finally, in section 5, we give a numerical assessment of the new method by applying it to the APSG reference wave function and show it as being superior to the diagonal partitioning applied previously.

2. Theory

We assume that the normalized zero-order wave function ψ satisfies the zero-order equation

$$\hat{H}^{(0)}|\psi\rangle = E^{(0)}|\psi\rangle \quad (2)$$

and we search the improvement to ψ and $E^{(0)}$ in an order-by-order expansion as

$$\Psi = \psi + \psi^{(1)} + \dots$$

and

$$E = E^{(0)} + E^{(1)} + E^{(2)} + \dots$$

where Ψ and E are the exact eigenstate and eigenenergy of the full Hamiltonian \hat{H} partitioned as

$$\hat{H} = \hat{H}^{(0)} + \hat{W}$$

To define a Fermi vacuum, let us distinguish a principal determinant in ψ , denoted by |HF⟩ [depending on the molecular orbitals, |HF⟩ may or may not be the Hartree–Fock (HF) determinant]

$$|\psi\rangle = c_{\text{HF}}|\text{HF}\rangle + \sum_{K \in V_R} c_K|K\rangle$$

and let us assume that c_{HF} is nonzero. Here and further on, notation $|K\rangle$, $|L\rangle$, etc. is used to indicate determinants excited with respect to |HF⟩. Occupied and virtual indices as well as the excitation level of determinants $|K\rangle$ will be also identified on the basis of the principal determinant |HF⟩. Set V_R collects indices of those excited determinants which have nonzero contribution to the reference function.

Provided that c_{HF} is nonzero, function ψ together with excited determinants $|K\rangle$ span the configuration space and form an overlapping basis. To construct a representation of the identity operator in terms of these vectors, we need to handle their overlap. This may be done by invoking any of the standard orthogonalization procedures which involve a numerical treatment of the overlap matrix. The overlap can be alternatively handled in an explicit manner if following a biorthogonal approach, due to the fact that the overlap matrix is invertible in a closed form. Two possible ways of a biorthogonal treatment are to

(a) Schmidt-orthogonalize $|K\rangle$'s to ψ as a first step, to obtain vectors

$$|K^\wedge\rangle = (1 - |\psi\rangle\langle\psi|)|K\rangle$$

In a second step, construct the reciprocal vectors to vectors $|K^\wedge\rangle$. This version of the theory is denoted pMCPT.

(b) Construct the reciprocal vector to the set formed by $|\psi\rangle$ and determinants $|K\rangle$. This version is denoted uMCPT. Alternatives a and b lead to a different definition of the projector corresponding to the one-dimensional model space, namely

(a) $\hat{O} = |\psi\rangle\langle\psi|$ is a symmetrical projector if Schmidt-orthogonalization is applied first

(b) $\hat{O} = |\psi\rangle\langle\tilde{\psi}|$ is a skew projector if the reciprocal set is constructed right away. A tilde is used for denoting reciprocal vectors, i.e., $\langle\tilde{L}|K\rangle = \delta_{LK}$

The sum $\hat{O} + \hat{P}$ is invariant to the choice of basis vectors; hence, a difference in the definition of \hat{O} results in a difference in \hat{P} as well. This is of importance since projectors \hat{O} and \hat{P} enter the definition of the zero-order Hamiltonian (eq 1). Consequently, the partitioning and the resulting PT series become different in the case of a and b. Before developing the PT treatment in the two cases, let us specify the Fockian, since it is common to both variants.

The Fockian \hat{F} enters the zero-order Hamiltonian projected appropriately by \hat{O} and \hat{P} to ensure fulfillment of the zero-

order Schrödinger—eq 2. We employ here a Fockian of the ordinary single reference form, constructed using the density matrix corresponding to the principal determinant. In the spin-orbital basis

$$\hat{F} = \sum_{ij} f_{ij} i^+ j^- = \sum_{ij} (h_{ij} + \sum_k^{\text{occ}} \langle ik||jk \rangle) i^+ j^-$$

with $\langle ik||jk \rangle$ standing for antisymmetrized two-electron integrals in the $\langle 12||12 \rangle$ convention. In accordance with the noncorrelated form of the Fockian, the zero-order energy of both variants is defined as

$$E^{(0)} = \langle \text{HF} | \hat{F} | \text{HF} \rangle$$

just like in ordinary single-reference MP theory.

Considering computational economy, it is obvious that the projection of \hat{F} into space \hat{P} , as shown in eq 1, is impractical, since the matrix of $\hat{H}^{(0)}$ is nondiagonal, with offdiagonal elements coupling subspaces of different excitation levels. In the actual calculations, the expression of eq 1 is simplified, as detailed in section 4.

2.1. pMCPT: Schmidt-Orthogonalization Prior to Reciprocal Set Construction. Schmidt-orthogonalization of determinant $|K\rangle$ to ψ produces

$$|K'\rangle = |K\rangle - c_K |\psi\rangle \quad (3)$$

Obviously, $|K'\rangle = |K\rangle$ for $K \notin V_R$. Vectors $|K'\rangle$ together with ψ form a basis in the configuration space. This basis is not orthogonal, however, as projected determinants may exhibit nonzero overlap among themselves. Reciprocal vectors to $|K'\rangle$ are given by¹¹

$$\langle \tilde{K}' | = \langle K | - \frac{c_K}{c_{\text{HF}}} \langle \text{HF} | \quad (4)$$

Again, $\langle \tilde{K}' | = \langle K |$ if $K \notin V_R$. Since the biorthogonal treatment affects only excited vectors, projector \hat{O} is symmetrical

$$\hat{O} = |\psi\rangle\langle\psi|$$

The energy up to first order is also given by the symmetrical expression

$$E^{(0)} + E^{(1)} = \langle \psi | \hat{H} | \psi \rangle = E_{\text{ref}} \quad (5)$$

Projector \hat{P} , expressed with excited determinants and their reciprocal counterparts, reads as

$$\hat{P} = \sum_K |K'\rangle\langle\tilde{K}'| \quad (6)$$

Note that in spite of \hat{P} looking like a skew-projector, it is an ordinary Hermitean projector, since $\hat{P} = 1 - \hat{O}$. Given the expressions of $E^{(0)}$, \hat{F} , \hat{O} and \hat{P} , the zero-order Hamiltonian is now well-defined by eq 1.

Imposing intermediate normalization for the wave function

$$\langle \psi | \Psi \rangle = 1 \quad (7)$$

implies that the first-order correction satisfies

$$\langle \psi | \psi^{(1)} \rangle = 0$$

giving rise to the expansion

$$|\psi^{(1)}\rangle = \sum_{K \in V_1} t_K |K'\rangle \quad (8)$$

Here, V_1 collects indices of those vectors which interact with $|\psi\rangle$ via the Hamiltonian, i.e., $\langle \tilde{K}' | \hat{H} | \psi \rangle \neq 0$. Set V_1 is of course much larger than V_R . It includes HF and elements of V_R in the general case, while it may be reduced if introducing approximations. Coefficients t_K are determined from the first-order equation

$$(\hat{H}^{(0)} - E^{(0)})|\psi^{(1)}\rangle = (E^{(1)} - \hat{W})|\psi\rangle \quad (9)$$

projected by $\langle \tilde{L}' | \in V_1$ to get

$$\sum_{K \in V_1} \langle \tilde{L}' | \hat{F} - E^{(0)} | K' \rangle t_K = -\langle \tilde{L}' | \hat{H} | \psi \rangle \quad (10)$$

In obtaining eq 10, the zero-order Hamiltonian (eq 1) was substituted on the left-hand side; the zero-order equation (eq 2) and $\langle \tilde{L}' | \psi \rangle = 0$ were applied on the right-hand side.

In the general case, the linear system of eq 10 determines the first-order wave function. Upon substituting eq 3 for $|K'\rangle$ and eq 4 for $\langle \tilde{L}' |$ one obtains

$$\begin{aligned} & \sum_{K \in V_1} \langle L | \hat{F} - E^{(0)} | K \rangle t_K - \langle L | \hat{F} - E^{(0)} | \psi \rangle \sum_{K \in V_1} c_K t_K \\ & - \frac{c_L}{c_{\text{HF}}} \sum_{K \in V_1} \langle \text{HF} | \hat{F} - E^{(0)} | K \rangle t_K + \frac{c_L}{c_{\text{HF}}} \langle \text{HF} | \hat{F} - E^{(0)} | \psi \rangle \sum_{K \in V_1} c_K t_K \\ & = -\langle L | \hat{H} | \psi \rangle + c_L \tilde{E}_{\text{ref}} \end{aligned} \quad (11)$$

where

$$\tilde{E}_{\text{ref}} = \langle \text{HF} | \hat{H} | \psi \rangle / c_{\text{HF}}$$

It is possible to simplify eq 11 if restricting ourselves to APSPG reference functions, which include exclusively doubly excited determinants expressed in the natural basis. This structure allows one to omit the fourth term on the left-hand side of eq 11. Furthermore, we restrict set V_1 to include only index of doubly excited determinants. This approximation eliminates the third term on the left-hand side of eq 11. Altogether, this means that reciprocal vector $\langle \tilde{L}' |$ can be substituted by $\langle L |$ on the left-hand side of eq 10, leading to the equations

$$\begin{aligned} & \sum_K^{2 \times \text{exc.}} \langle L | \hat{F} - E^{(0)} | K \rangle t_K - \langle L | \hat{F} - E^{(0)} | \psi \rangle \sum_K^{2 \times \text{exc.}} c_K t_K = \\ & -\langle L | \hat{H} | \psi \rangle + c_L \tilde{E}_{\text{ref}} \end{aligned} \quad (12)$$

The second-order equation

$$\hat{H}^{(0)}|\psi^{(2)}\rangle + \hat{W}|\psi^{(1)}\rangle = E^{(0)}|\psi^{(2)}\rangle + E^{(1)}|\psi^{(1)}\rangle + E^{(2)}|\psi\rangle \quad (13)$$

projected by $\langle \psi |$ gives the second-order energy

$$E^{(2)} = \langle \psi | \hat{H} | \psi^{(1)} \rangle = \sum_K^{2 \times \text{exc.}} \langle \psi | \hat{H} - c_K E_{\text{ref}} | K \rangle t_K \quad (14)$$

having utilized the fact that $\langle\psi|$ is an eigenfunction of $\hat{H}^{(0)}$ from the left, normalization condition (eq 7), expansion (eq 8), eqs 3 and 5. Equations 12 and 14 are the working equations of the method MP-pMCPT(APSG) presented in the applications, where an APSG reference function is adopted.

2.2. uMCPT: Reciprocal Set Construction without Schmidt-Orthogonalization. Reciprocal vectors to the set formed by $|\psi\rangle$ and $|K\rangle$'s can be given by

$$\langle\tilde{\psi}| = \frac{1}{c_{\text{HF}}}\langle\text{HF}| \quad (15)$$

and

$$\langle\tilde{K}| = \langle K| - \frac{c_K}{c_{\text{HF}}}\langle\text{HF}|$$

With the use of the above vectors, one can put down skew-projector \hat{O} in the form

$$\hat{O} = |\psi\rangle\langle\tilde{\psi}|$$

The sum of zero and first-order energies is also evaluated on the basis of the nonsymmetrical expression

$$E^{(0)} + E^{(1)} = \langle\tilde{\psi}|\hat{H}|\psi\rangle = \tilde{E}_{\text{ref}}$$

This energy expression is equivalent to the symmetric form E_{ref} of eq 5 in the case where coefficients in the expansion of ψ are determined from diagonalization of \hat{H} in a subspace of the configuration space. This holds true for an MCSCF wave functions or functions produced by single- or multi-reference CI procedures but not for the APSG wave function considered in the present applications. A skew-projector orthogonal and complementary to \hat{O} is written as

$$\hat{P} = \sum_K |K\rangle\langle\tilde{K}| \quad (16)$$

With the above \hat{O} and \hat{P} definition and $E^{(0)}$ and \hat{F} remaining unaltered, the zero order Hamiltonian of uMCPT is again defined by eq 1.

A suitable form of the intermediate normalization condition in this version of the theory is

$$\langle\tilde{\psi}|\Psi\rangle = 1 \quad (17)$$

Consequently, the first-order wave function should satisfy

$$\langle\tilde{\psi}|\psi^{(1)}\rangle = 0$$

Hence, in terms of vectors $|K\rangle$, it can be expanded as

$$|\psi^{(1)}\rangle = \sum_{K \in V_1} t_K |K\rangle \quad (18)$$

In this formulation, HF is missing from V_1 , due to the normalization (eq 17). Coefficients t_K are determined from the first-order eq 9, projected by $\langle\tilde{L}| \in V_1$ to get

$$\sum_{K \in V_1} \langle\tilde{L}|\hat{F} - E^{(0)}|K\rangle t_K = -\langle\tilde{L}|\hat{H}|\psi\rangle \quad (19)$$

In obtaining eq 19, the form of the zero-order Hamiltonian (eq 1) was applied, as well as the zero-order eq 2 and $\langle\tilde{L}|\psi\rangle = 0$. The general form of the equations determining function $\psi^{(1)}$ in this variant of the theory is eq 19.

Considering the approximation where V_1 is restricted to doubly excited indices, term $-c_L\langle\text{HF}|\hat{F} - E^{(0)}|K\rangle t_K / c_{\text{HF}}$ stemming from the overlap of $\langle\tilde{L}|$ with $|\psi\rangle$ can be omitted on the left-hand side of eq 19, leading to

$$\sum_K^{2 \times \text{exc.}} \langle\tilde{L}|\hat{F} - E^{(0)}|K\rangle t_K = -\langle\tilde{L}|\hat{H}|\psi\rangle + c_L \tilde{E}_{\text{ref}} \quad (20)$$

The second-order eq 13 projected by $\langle\tilde{\psi}|$ gives the second-order energy

$$E^{(2)} = \langle\tilde{\psi}|\hat{H}|\psi^{(1)}\rangle = \frac{1}{c_{\text{HF}}} \sum_K^{2 \times \text{exc.}} \langle\text{HF}|\hat{H}|K\rangle t_K \quad (21)$$

having utilized that $\langle\tilde{\psi}|$ is an eigenfunction of $\hat{H}^{(0)}$ from the left, normalization condition (eq 17), eq 15, and expansion (eq 18). Equations 20 and 21 are the working equations of the method denoted MP-uMCPT(APSG) in the applications, where an APSG reference function is adopted.

3. Size-Consistency

Among previous versions of the theory, where the zero-order Hamiltonian was assumed diagonal, uMCPT was shown to provide size-consistent correction at second order, if energy denominators were composed of one-particle indexed quantities.¹¹ We investigate here whether this property subsists in MP-uMCPT and find that canonical orbitals in the single-reference sense ensure a second-order energy behaving well in this respect. For noncanonical orbitals, deletion of the occupied-virtual block of the Fockian in the definition of $\hat{H}^{(0)}$ is necessary to obtain this behavior.

By size consistency, we understand the criterion of obtaining the energy as a sum of subsystem energies in the case where subsystems do not interact. To study this, let us suppose that the reference function is behaving well; i.e., it is given as a product [antisymmetrization being immaterial for noninteracting subsystems¹⁴] of noninteracting partner's reference functions

$$|\psi\rangle = |\psi_A \psi_B\rangle$$

where index A and B label the subsystems. As a consequence, the principal determinant is also given as the product

$$|\text{HF}\rangle = |\text{HF}_A \text{HF}_B\rangle$$

appearing in the expansion of $|\psi\rangle$ with weight $c_{\text{HF}}^A c_{\text{HF}}^B$; hence, the reciprocal vector $\langle\tilde{\psi}|$ reads

$$\langle\tilde{\psi}| = \langle\text{HF}_A \text{HF}_B| / (c_{\text{HF}}^A c_{\text{HF}}^B)$$

Since both the total Hamiltonian and the Fockian are given as a sum over noninteracting systems, the reference energy

$$\tilde{E}_{\text{ref}} = \tilde{E}_{\text{ref}}^A + \tilde{E}_{\text{ref}}^B$$

and the zero-order energy

$$E^{(0)} = E_A^{(0)} + E_B^{(0)}$$

separate for terms corresponding to individual subsystems.

Determinants appearing in the expansion of $|\psi^{(1)}\rangle$ can be classified as doubly excited on system A, doubly excited on system B, or singly excited both on system A and B, giving rise to the form

$$|\psi^{(1)}\rangle = \sum_K^A t_{KHF}^{AB} |K_A HF_B\rangle + \sum_K^B t_{HF_K}^{AB} |HF_A K_B\rangle + \sum_K^A \sum_I^B t_{KI}^{AB} |K_A I_B\rangle \quad (22)$$

with self-explanatory notations. The above expansion substituted into the coefficients' equation (eq 20), we have to consider two distinct cases: (i) index L refers to a determinant doubly excited on one subsystem, say A, or (ii) index L belongs to a determinant singly excited on both subsystems. In case i, $\langle L|$ can be written as

$$\langle L| = \langle L_A HF_B|$$

and by trivial derivation, one arrives at the coefficient equation

$$\sum_K^A \langle L_A | \hat{F}_A - E_A^{(0)} | K_A \rangle t_{KHF}^{AB} + \sum_I^B \langle HF_B | \hat{F}_B | I_B \rangle t_{LI}^{AB} = -\langle L_A | \hat{H}_A - \tilde{E}_{\text{ref}}^A | \psi_A \rangle c_{HF}^B$$

This equation should contain solely quantities belonging to subsystem A, which does not hold because of the second term on the left-hand side. (Coefficient c_{HF}^B does not do any harm; in fact, it has a proper role as seen in eq 23.) In the study of case ii, it is seen that $\langle L|$ adopts the form

$$\langle L| = \langle L_A J_B|$$

and the coefficient equation is found to be

$$\langle J_B | \hat{F}_B | HF_B \rangle t_{LHF}^{AB} + \langle L_A | \hat{F}_A | HF_A \rangle t_{HFJ}^{AB} + \sum_K^A \langle L_A | \hat{F}_A - E_A^{(0)} | K_A \rangle t_{KJ}^{AB} + \sum_I^B \langle J_B | \hat{F}_B - E_B^{(0)} | I_B \rangle t_{LI}^{AB} = -\langle L_A | \hat{H}_A - \tilde{E}_{\text{ref}}^A | \psi_A \rangle c_J^B - \langle I_B | \hat{H}_B - \tilde{E}_{\text{ref}}^B | \psi_B \rangle c_L^A$$

Due to A and B being noninteracting, coefficients of the type t_{LI}^{AB} do not contribute to the second-order energy. The above equation—which corresponds to these rows—is therefore not important provided it is not coupled to columns corresponding to local excitations, e.g., $|L_A HF_B\rangle$. Unfortunately, the first two terms on the left-hand side are just consistency-spoiling coupling terms. When the two cases are summarized, the coefficient matrix on the left-hand side of eq 20 can be depicted as shown in Figure 1.

Substituting expansion 22 into the second-order energy formula, one obtains

$$E^{(2)} = \sum_K^A \langle \tilde{\psi}_A | \hat{H}_A | K_A \rangle t_{KHF}^{AB} \frac{1}{c_{HF}^B} + \{\text{A} \leftrightarrow \text{B exchanged}\} \quad (23)$$

indicating that size consistency would hold if the equation determining t_{KHF}^{AB}/c_{HF}^B would be the same as the equation for t_K^A , when computed alone. This is spoiled by the coupling emerging in the blocks dotted in Figure 1. Nonzero elements of these blocks are solely occupied-virtual matrix elements of the Fockian and are zero only if the orbitals are canonical in the single-reference sense. In general, it certainly does not hold for multireference applications. To restore size consistency in such a case, one can modify the partitioning by allowing nonzero elements only in the occupied–occupied and virtual–virtual block of the Fockian.

4. Properties of MP-MCPT and Survey of Related Theories

Several MR extensions of MP theory are related to the MP-MCPT scheme detailed above. A characteristic feature unique to the MCPT framework is the biorthogonal treatment of the overlap among basis vectors. This is in contrast to the approach introduced by Wolinsky and co-workers^{15,16} where internally contracted excited vectors are considered as basis vectors and Schmidt-orthogonalization is applied to keep subspaces of different excitation levels orthogonal to each other. Vectors belonging to the same excited subspace can be orthogonalized either by Löwdin's symmetrical¹⁷ or canonical scheme^{18,19} or by the Gram–Schmidt procedure.²⁰ Diagonalization of the overlap matrix can become a bottleneck of this approach, which induces the application of a partially contracted and partially uncontracted basis.^{17,21} To avoid the overlap problem, Murphy and Messmer suggested the use of totally uncontracted configuration state functions (CSFs) as basis vectors in the excited space.^{22,23} This theory has to cope with an increased dimension of the linear system of equations to solve in return. Both the approaches of Murphy and Messmer and the MRMP method introduced by Hirao et al.^{24,25} assume the existence of a set of multiconfigurational basis vectors orthogonal and noninter-

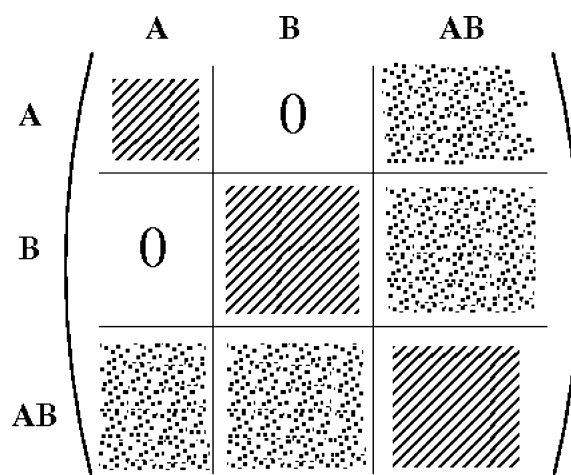


Figure 1. Block structure of the coefficient matrix of the first-order equation for noninteracting systems A and B.

acting through \hat{H} with the reference function (e.g., excited CAS vectors). Explicit construction of these multiconfigurational basis vectors becomes necessary only beyond third order in energy, which was never investigated with these theories to the best of our knowledge. Within the MRMP framework, McDouall and Robinson have conducted extensive research in the line of lifting orbital optimization problems as well as reducing the size of the model space, see ref 26 and references therein.

Specific treatment of overlap among excited basis vectors is of course irrelevant as far as the zero-order Hamiltonian is of the form eq 1 and $E^{(0)}$ and \hat{O} are defined alike. Most methods however do not apply the zero-order operator (eq 1) as it is. In their pioneering paper, Wolinsky et al.¹⁵ suggested decoupling interactions at the zero order using the Hamiltonian

$$\hat{H}^{(0)} = E^{(0)}\hat{O} + \hat{P}_S\hat{F}\hat{P}_S + \hat{P}_D\hat{F}\hat{P}_D + \dots \quad (24)$$

to break down the dimension of the inversion problem for smaller sub-blocks. Here, \hat{P}_S , \hat{P}_D , etc. refer to singly, doubly, etc. excited subspaces. With such a zero-order Hamiltonian, the definition of \hat{P}_S , \hat{P}_D , etc. clearly becomes of importance and affects the behavior of the PT series. Several different decoupling schemes have been investigated over time,^{18,20} while Celani and Werner reported second-order energies with the nondecoupled zero order of eq 1.²¹ It was also shown that increasing the block-diagonal character of $\hat{H}^{(0)}$ reduces the size-consistency error of individual energy corrections.^{20,27}

The MP-MCPT framework avoids the overlap problems present in internally contracted theories by adopting a determinant-based description and a biorthogonal treatment. At the same time, the dimension of the linear system of equations is kept at a manageable size by a decoupling of the type given in eq 24. In fact, restricting expansion of the first-order function to doubly excited determinants means that the zero-order Hamiltonian of MP-MCPT reads

$$\hat{H}^{(0)} = E^{(0)}\hat{O} + \hat{P}_D\hat{F}\hat{P}_D \quad (25)$$

where \hat{P}_D is either of the form eq 6 or eq 16, with summation index K restricted to doubly excited determinants. This zero-order Hamiltonian is of course unfitted for calculating energies beyond third order. Even third-order results are omitted from the present study, where we intentionally aim to capture a significant portion of the dynamical correlation energy at the lowest order of a simple perturbation scheme. The error committed by decoupling of eq 25 as compared to eq 1 is expected to be negligible at order 2. At the same time, decoupling eq 25 means that the coefficient matrix on the left-hand side of eq 19 is of *exactly* the same form as the matrix appearing in single-reference MP calculations performed on a localized basis.^{28–30} The inversion of this matrix is the rate determining step of MP-MCPT. Since the Fockian is a one-body operator, the structure of the coefficient matrix of the linear system of equations is comfortably sparse and easily invertible by iterative algorithms.^{31,32} In the MP-pMCPT variant of the theory, a correction term on the left-hand side of eq 10 makes a difference with the coefficient matrix of single-reference MP theory. This

correction affects those columns which correspond to the determinants present in the expansion of ψ but does not alter the size of the matrix.

The definition of the Fockian as well as the zero-order energy $E^{(0)}$ is an important question in MR MP theories, related to the sensitivity to intruder states. Most MP extensions use the generalized Fockian³³ built with the correlated one-body density matrix of the reference function and define the zero-order reference energy as the expectation value $\langle\psi|\hat{F}|\psi\rangle$. At odds with these, the density matrix of the principal determinant is used to construct the Fockian in MP-MCPT and we take $\langle\text{HF}|\hat{F}|\text{HF}\rangle$ as zero-order energy, both being the same constructions as in single-reference MP. Our choice is motivated partly by computational simplicity and partly by previous numerical experiences,¹¹ indicating a negligible difference in second-order results between using the uncorrelated or generalized Fockian. In fact, a generalized Fockian fits better to a multiconfiguration framework, and it is preferred particularly if orbital invariance of the theory is desirable. In our approach, however, a principal determinant is pinpointed at the stage of defining reciprocal vectors. This inhibits invariance to any orbital rotations and enhances the single-reference character of the theory, making it rather pointless to apply a generalized Fockian. Defining the ground state zero-order energy as in single-reference MP theory appears particularly dangerous due to the well-known quasi-degeneracy problem upon bond-dissociation. This fear, however, is just slightly justified according to the numerical experiences presented in section 5. On the other hand, working with a spectral representation of the Fockian built with CASSCF orbitals and orbital energies has been found to give a poor description of multireference problems if the reference function is a single configuration state function.³⁴

As already alluded to, MP-MCPT is not invariant to orbital rotations that may leave the reference function unaffected. This is undesirable, but not unique among MR MP theories; e.g., assumption of a diagonal form of $\hat{H}^{(0)}$ destroys the invariance.^{8,10,11,35} In the case of MP-MCPT, orbital non-invariance stems from the biorthogonal treatment and has the further consequence that MP-MCPT is not invariant to the choice of principal determinant either. This suggests that MP-MCPT is safely applicable only in the case where one of the determinants stands out in the expansion of the reference function, in terms of coefficient squared. The dissociation of the nitrogen molecule, where several determinants become equally weighted at the end of the process, is one test of this feature. As shown in section 5, performance of MP-MCPT is surprisingly acceptable in this example apart from the slight breakdown of the curve. In contrast to the nitrogen dissociation example, serious qualitative failure is in fact observed when the principal role is handed over from one determinant to another during the process studied. These are cases where MP-MCPT definitely should not be applied as it is. Averaging over principal determinants has been shown to be a possible cure to this problem.³⁶

Choosing a suitable two-body zero-order Hamiltonian satisfying eq 2 instead of definition of eq 1 is certainly superior to any MP extension discussed here, and such theories were shown to produce excellent results,^{37–40} at the

price of coping with a more tedious task when obtaining the first-order coefficients. The present theory—being an uncomplicated version even among MP theories assuming a one-body zero-order Hamiltonian—obviously cannot compete with these methods either in accuracy or in desirable properties like size consistency or orbital invariance. On the other hand, we do observe an improvement in the numerical performance as compared to considering a diagonal zero-order Hamiltonian within the MCPT framework, suggested previously,^{10,11,41} although in some cases the improvement in total energies may be rather small.

5. Assessments

We assessed the present MP-MCPT methods by adopting the APSG wave function expressed in the natural orbital basis as a reference. The APSG function can be written as the products of ground and pair-excited geminal functions as follows:

$$|\psi\rangle \equiv |\psi^{\text{APSG}}\rangle = c_{\text{HF}} \prod_i^{\text{geminal}} \left(1 + \sum_{a \in \mathbf{S}(i)} \frac{c_i^a}{c_{\text{HF}}} \hat{T}_{i a a \beta}^a \right) |\text{HF}\rangle$$

where $\mathbf{S}(i)$ is the set of the unoccupied orbitals of the geminal subset which has an occupied orbital i . We restricted the expansion of the first-order wave function within doubly excited determinants from $|\text{HF}\rangle$, as discussed previously.

We selected the H₂O (water), HF (hydrogen fluoride), N₂ (nitrogen), and F₂ (fluorine) molecules as test systems and obtained potential energy curves for the bond dissociations. As a comparison, we present APSG, MP2, multireference MP2 (MRMP2), and a PT designed for the APSG wave function by Rosta and Surján (APSG-PT).^{40,42} In addition, we also computed the equilibrium geometries of the diatomic molecules and calculated vibrational frequencies by the finite difference method. During the latter, we first determined equilibrium distances R_e up to the order of 0.1 pm and evaluated the frequencies from three points, namely, R_e and $R_e \pm 0.5$ pm structures. All calculations were performed with the 6-311G** basis set.⁴³ The APSG geminal subsets were defined to give six orbitals for each bonding geminal and three orbitals for each lone-pair geminal, around the equilibrium structure.

5.1. Dissociation Curves. We first calculated potential energy curves for two types of bond-breaking processes of the H₂O molecule: (i) a heterogeneous one-bond dissociation, with the other bond distance fixed to 95 pm, and (ii) a homogeneous two-bond dissociation. In both processes, the bond angle was fixed to 104.5°. The reference function underlying the MRMP2 calculation was a CASSCF wave function with four active electrons on four active orbitals, CASSCF(4,4) shortly.

Figure 2 shows the potential energy curves for one-bond dissociation of H₂O. The APSG curve is much worse in absolute energy than MRMP2. However, APSG can produce a qualitatively nice dissociation curve: nonparallelity error (NPE) with respect to MRMP2 is 0.0160 hartree. The single-reference perturbation approach (MP2) starts to diverge at about the 300 pm structure. Around equilibrium distance, both MP-pMCPT and MP-uMCPT are much improved from

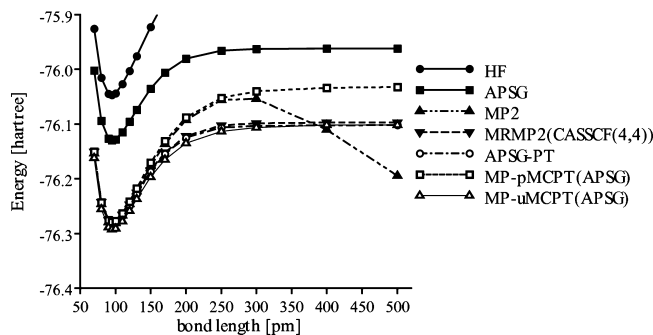


Figure 2. Potential energy curves for the heterogeneous one-bond dissociation of the H₂O molecule. The other O—H bond length is fixed to 95 pm and the bond angle is fixed to 104.5°.

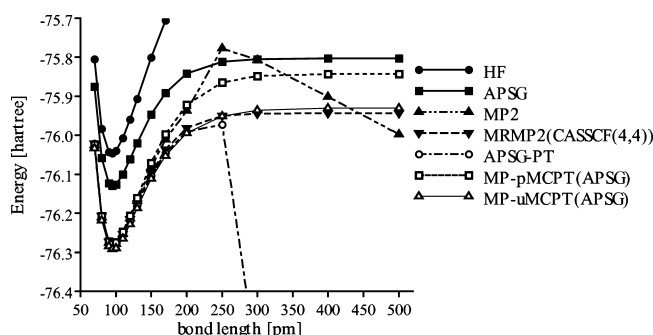


Figure 3. Potential energy curves for the homogeneous two-bond dissociation of the H₂O molecule. The bond angle is fixed to 104.5°.

APSG in absolute energy, due to the consideration of dynamical correlation. As the bond length gets large, however, the two curves behave differently. The curve by MP-pMCPT becomes similar to the MP2 one up to 250 pm and levels out; hence, the dissociation energy is overestimated compared to MRMP2. On the other hand, MP-uMCPT reproduces the shape by MRMP2 or APSG-PT up to the dissociation limit. This may be attributed to the quasi size consistency of MP-uMCPT.

Figure 3 shows the potential energy curves for two-bond homogeneous dissociation of the H₂O molecule. Although this sort of dissociation requires at least four-electron four-orbital active space, APSG still gives a qualitatively nice curve. APSG-PT cannot produce a correct dissociation curve shape in this case. On the other hand, MP-pMCPT and MP-uMCPT nicely level out with increasing bond length. The MP-pMCPT curve again follows MP2 up to 200 pm and overestimates the dissociation energy as compared to the MRMP2 result. The MP-uMCPT produces a potential curve similar to MRMP2 even for this multiple bond dissociation example.

Next, we assessed the dissociation potential energy surface for the bond-breaking process of the HF molecule, shown in Figure 4. In this system, the full-configuration interaction (FCI) results were obtained around equilibrium and dissociated structures by utilizing the sparse FCI algorithm of Rolik et al.⁴⁴ The behavior of the curves resembles that of Figure 2. In particular, MP-uMCPT reproduces the MRMP2 curve well, while MP-pMCPT follows the MP2 curve up to 250

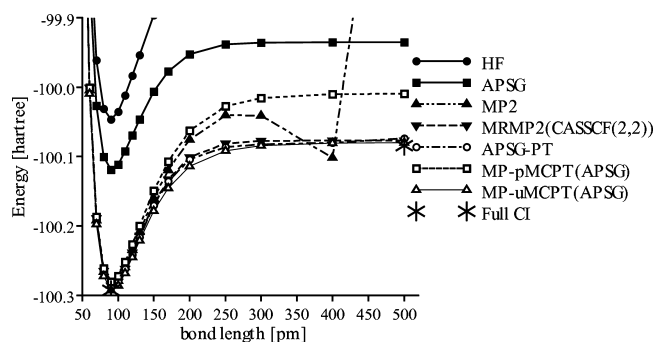


Figure 4. Potential energy curves for the dissociation of the HF molecule.

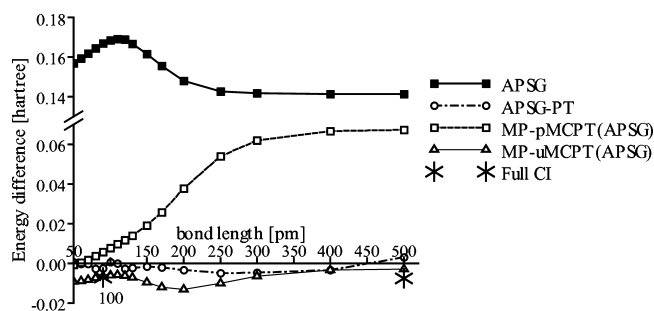


Figure 5. Energy difference from the MRMP2 results for the dissociation of the HF molecule.

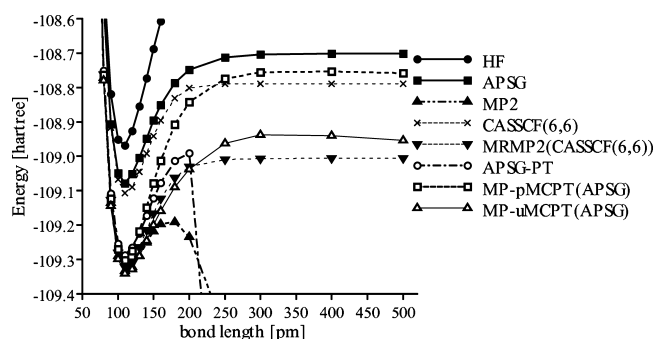


Figure 6. Potential energy curves for the dissociation of the N_2 molecule.

pm. Since the energy errors of MRMP2 with respect to FCI are comparable (0.0071 and 0.0075 hartree at 90 and 500 pm bond length, respectively), the energy difference from MRMP2 is a good indicator to assess the accuracy of the methods. These data are shown in Figure 5. Around equilibrium distance, the APSPG energy error is larger than at the end of the process, due to the lack of dynamical correlation. The errors of MP-pMCPT and MP-uMCPT around equilibrium geometry are improved to less than 0.01 hartree by taking dynamical correlation into account. While the error of MP-pMCPT becomes large as the bond is stretched, MP-uMCPT remains fairly constant: NPEs of MP-pMCPT and MP-uMCPT are 0.0671 and 0.0102 hartree. The latter is comparable to the 0.0079 hartree error of APSPG-PT.

Further, we assessed potential energy curves for the triple-bond-breaking process of the N_2 molecule, shown in Figure 6. The MRMP2 calculations for the N_2 molecule were based on a CASSCF(6,6) wave function as a reference. In this example, APSPG-PT as well as MP2 diverge, as expected.

Table 1. Calculated Equilibrium Distances (R_e), Harmonic Vibrational Frequencies (f), and Dissociation Energies (D_e) of the HF Molecule Adopting the 6-311G** Basis Set

method	R_e [pm]	f [cm^{-1}]	D_e [eV]
HF	89.6	4496	22.14
APSPG	91.0	4223	4.997
MP2	91.2	4247	
MRMP2(CASSCF(2,2))	91.9	4143	5.696
APSPG-PT	91.8	4038	5.842
MP-pMCPT(APSPG)	91.0	4280	7.368
MP-uMCPT(APSPG)	91.6	4160	5.789
full CI	91.3	4213	5.679

Table 2. Calculated Equilibrium Distances (R_e), Harmonic Vibrational Frequencies (f), and Dissociation Energies (D_e) of the F_2 Molecule Adopting the 6-311G** Basis Set

method	R_e [pm]	f [cm^{-1}]	D_e [eV]
HF	133.1	1209	9.347
APSPG	153.2	521	0.475
MP2	141.1	914	
MRMP2(CASSCF(2,2))	144.8	759	1.233
APSPG-PT	146.1	711	1.068
MP-pMCPT(APSPG)	136.8	1087	4.089
MP-uMCPT(APSPG)	148.0	678	1.538
exptl. ^a	141.2	917	1.602

^a Ref 45.

The MP-pMCPT and MP-uMCPT methods give qualitatively good dissociation profiles even for this triple-bond breaking, although slight bumps can be seen between the equilibrium and dissociated structures. It is to be noted here that the APSPG reference wave function underlying MP-MCPTs is poorer than CASSCF(6,6) used for MRMP2, since sextuply excited determinants appear as products of two-electron excitations in APSPG. The imperfection of APSPG to describe triple bond breaking as compared with CASSCF may be credited for the breakdown of MP-MCPT dissociation curves.

5.2. Parameters at Equilibrium Geometry. Next, we calculated parameters at an equilibrium geometry of diatomic molecules, i.e., equilibrium bond length (R_e), harmonic frequencies (f), and dissociation energies (D_e). Dissociation energy is evaluated as the energy difference between the equilibrium and 500 pm structures.

In Table 1, we summarize the parameters of the HF molecule. Equilibrium bond distance as calculated by either of the present MP-MCPTs agrees with FCI within 0.3 pm. This is better than the R_e obtained by either MRMP2 or APSPG-PT. The MP-pMCPT frequency is larger than f calculated by MP-uMCPT or MRMP2, which relates to the overestimation of the dissociation energy in MP-pMCPT. Both R_e and f are remarkably well estimated by APSPG in this system.

The situation becomes different in the F_2 molecule, which has a much shallower potential than HF. Table 2 shows the parameters of F_2 . For comparison, experimental data from ref 45 are also indicated. Compared to experimental values, APSPG overestimates R_e by more than 10 pm and underestimates D_e by 70%, which is also reflected in the underestimation of f . As a contrast to this, MP-pMCPT underestimates R_e by about 5 pm, overestimates D_e by more than 200%, and consequently also overestimates the harmonic

Table 3. Calculated Equilibrium Distances (R_e), Harmonic Vibrational Frequencies (f), and Dissociation Energies (D_e) of the N_2 Molecule Adopting the 6-311G** Basis Set

method	R_e [pm]	f [cm^{-1}]	D_e [eV]
HF	107.0	2732	30.91
APSG	109.3	2455	10.11
MP2	111.9	2178	
CASSCF(6,6)	110.7	2349	8.646
MRMP2(CASSCF(6,6))	111.1	2295	8.597
MP-pMCPT(APSG)	109.3	2507	14.78
MP-uMCPT(APSG)	111.7	2231	10.53
exptl. ^a	109.8	2359	9.759

^a Ref 45.

frequency. On the other hand, MP-uMCPT gives reasonable results: D_e is much improved from APSG, and R_e and f agree with those by APSG-PT or MRMP2 tolerably.

Finally, the parameters of N_2 are summarized in Table 3 and compared to experimental data from ref 45. The equilibrium bond length and the harmonic frequency are well reproduced within 2 pm and 150 cm^{-1} except for HF and MP2. Overestimation of R_e and slight underestimation of f is given by MP-uMCPT, showing resemblance to MRMP2 results. However, MP-uMCPT overestimates D_e , which is contrary to MRMP2. The overshooting of D_e is larger by MP-pMCPT, about 150%.

6. Concluding Remarks

Two simple extensions of single-reference MP theory to the multireference case were presented at the second order. The theories are strongly reminiscent of the single-reference MP2 procedure, particularly in what concerns the coefficient matrix of the linear system of equations determining the first-order wave function. Considering this equation, the present MR extensions practically affect only the inhomogeneous term, i.e., the right-hand side of the first-order equation. Numerical implementation of the theories is straightforward on the basis of an existing single-reference code adapted to a localized basis. Computational requirements of the approaches agree with single-reference MP2 calculation on a localized basis.

Among previous multireference MP theories, MP-MCPT shows the most similarity with multireference PT methods which apply a Fockian appropriately multiplied by Hilbert-space projectors to define the zero-order Hamiltonian. The novelty of the present scheme lies in the biorthogonal treatment of the overlap among basis vectors in the configuration space.

The simplicity of MP-MCPT methods is counterweighted by their failure to show desirable properties like orbital or principal determinant invariance. Size consistency is achievable only in MP-uMCPT, if assuming a block-diagonal form of the Fockian. Numerical assessment shows that in spite of their simplicity, the range of applicability does cover problems of significant multireference character, like the bond breaking process. Properties of equilibrium structures are also well estimated by MP-uMCPT.

Acknowledgment. The authors are grateful to Z. Rolik (Budapest) for performing full-CI calculations. This work

was supported by the Hungarian Research fund OTKA NI-67702, K-81590, K-81588, and by a Grant-in-Aid for Young Scientists (B) "KAKENHI 22750016" from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). One of the authors (M.K.) is indebted to a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science (JSPS).

References

- (1) Hinze, J. *J. Chem. Phys.* **1973**, *59*, 6424.
- (2) Roos, B. O. In *Advances in Chemical Physics*; Wiley & Sons Ltd: New York, 1987; Vol. 69, pp 339–445.
- (3) Bobrowicz, F. W.; Goddard, W. A., III. The Self-Consistent Field Equations for Generalized Valence Bond and Open-Shell Hartree-Fock Wave Functions. In *Methods of Electronic Structure Theory*; Schaefer, H. F., III, Ed.; Plenum: New York, 1977; p 79.
- (4) Hurley, A. C.; Lennard-Jones, J.; Pople, J. A. *Proc. R. Soc., London* **1953**, *A220*, 446.
- (5) Kutzelnigg, W. *J. Chem. Phys.* **1964**, *40*, 3640.
- (6) Kapuy, E. *J. Chem. Phys.* **1966**, *44*, 956.
- (7) Surján, P. R. *Top. Curr. Chem.* **1999**, *203*, 63.
- (8) Davidson, E. R.; Bender, C. *Chem. Phys. Lett.* **1978**, *59*, 369–374.
- (9) Kozłowski, P. M.; Davidson, E. R. *Chem. Phys. Lett.* **1994**, *222*, 615–620.
- (10) Rolik, Z.; Szabados, Á.; Surján, P. R. *J. Chem. Phys.* **2003**, *119*, 1922.
- (11) Szabados, Á.; Rolik, Z.; Tóth, G.; Surján, P. R. *J. Chem. Phys.* **2005**, *122*, 114104.
- (12) Rassolov, V. A.; Xu, F.; Garashchuk, S. *J. Chem. Phys.* **2004**, *120*, 10385–10394.
- (13) Rassolov, V. A.; Xu, F. *J. Chem. Phys.* **2007**, *127*, 044104.
- (14) Mayer, I. *Simple Theorems, Proofs, and Derivations in Quantum Chemistry*; Kluwer: New York, 2003; p 102.
- (15) Wolinski, K.; Sellers, H.; Pulay, P. *Chem. Phys. Lett.* **1987**, *140*, 225.
- (16) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **1989**, *90*, 3647.
- (17) Werner, H.-J. *Mol. Phys.* **1996**, *89*, 645–661.
- (18) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483.
- (19) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218.
- (20) van Dam, H. J. J.; van Lenthe, J. H. *Mol. Phys.* **1998**, *93*, 431–439.
- (21) Celani, P.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 5546.
- (22) Murphy, R. B.; Messmer, R. P. *Chem. Phys. Lett.* **1991**, *183*, 443.
- (23) Murphy, R. B.; Messmer, R. P. *J. Chem. Phys.* **1992**, *97*, 4170.
- (24) Hirao, K. *Chem. Phys. Lett.* **1993**, *201*, 59.
- (25) Choe, Y.; Witek, H. A.; Finley, J. P.; Hirao, K. *J. Chem. Phys.* **2001**, *114*, 3913.
- (26) Robinson, D.; McDouall, J. J. W. *J. Phys. Chem. A* **2007**, *111*, 9815.

- (27) van Dam, H.; van Lenthe, J.; Ruttink, P. *Int. J. Quantum Chem.* **1999**, *72*, 549–558.
- (28) Pulay, P.; Saebø, S. *Theor. Chim. Acta* **1986**, *69*, 357.
- (29) Saebø, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86*, 914.
- (30) Schütz, M.; Hetzer, G.; Werner, H.-J. *J. Chem. Phys.* **1999**, *111*, 5691–5705.
- (31) Pissanetzky, S. *Sparse Matrix Technology*; Academic Press: London, 1984.
- (32) Tuminaro, R. S.; Shadid, J. N.; Heroux, M. *Aztec: A massively parallel iterative solver library for solving sparse linear systems*, ver. 2.1; Sandia Corporation: Albuquerque, NM, 1999.
- (33) McWeeny, R. *Methods of Molecular Quantum Mechanics*; Academic: London, 1989.
- (34) Chen, F. *J. Chem. Theory Comput.* **2009**, *5*, 931.
- (35) Pariser, O.; Ellinger, Y. *Chem. Phys.* **1996**, *205*, 323–349.
- (36) Szabados, Á.; Surján, P. R. In *Progress in Theoretical Chemistry and Physics*; Springer: New York, 2009; pp 257–269.
- (37) Dyal, K. *J. Chem. Phys.* **1995**, *102*, 4909.
- (38) Mahapatra, U. S.; Datta, B.; Mukherjee, D. *Chem. Phys. Lett.* **1999**, *299*, 42–50.
- (39) Chattopadhyay, S.; Mahapatra, U. S.; Mukherjee, D. *J. Chem. Phys.* **1999**, *111*, 3820–3830.
- (40) Rosta, E.; Surján, P. R. *J. Chem. Phys.* **2002**, *116*, 878–890.
- (41) Surján, P. R.; Rolik, Z.; Szabados, Á.; Kóhalmi, D. *Ann. Phys. (Leipzig)* **2004**, *13*, 223–231.
- (42) Rosta, E.; Surján, P. R. *Int. J. Quantum Chem.* **2000**, *80*, 96.
- (43) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.
- (44) Rolik, Z.; Szabados, Á.; Surján, P. R. *J. Chem. Phys.* **2008**, *128*, 144101.
- (45) Huber, K.; Herzberg, G. *Molecular Spectra and Molecular Structure 4. Constants of Diatomic Molecules*; Van Nostrand: Princeton, NJ, 1979.

CT1001939

Eigenspace Update for Molecular Geometry Optimization in Nonredundant Internal Coordinate

Wenkel Liang,[†] Haitao Wang,[‡] Jane Hung,[†] Xiaosong Li,^{*,†} and Michael J. Frisch[§]

Department of Chemistry, University of Washington, Seattle, Washington 98195, The Computer Application Technology Key Lab of Yunnan Province, Kunming University of Science and Technology, Kunming, Yunnan, China 650093, Gaussian Inc., 340 Quinnipiac St, Bldg 40, Wallingford, Connecticut 06492

Received April 22, 2010

Abstract: An eigenspace update method is introduced in this article for molecular geometry optimization. This approach is used to obtain the nonredundant internal coordinate space and diagonalize the Hessian matrix. A select set of large molecules is tested and compared with the conventional method of direct diagonalization in redundant space. While all methods considered herein take on similar optimization pathways for most molecules tested, the eigenspace update algorithm becomes much more computationally efficient with increasing size of the molecular system. A factor of 3 speed-up in overall computational cost is observed in geometry optimization of the 25-alanine chain molecule. The contributing factors to the computational savings are the reduction to the much smaller nonredundant coordinate space and the $O(N^2)$ scaling of the algorithm.

I. Introduction

Molecular geometry optimization underlies all computational chemistry investigations by providing characteristic stationary point structures on potential energy surfaces (PESs).^{1,2} The most widely used geometry optimization method is the so-called quasi-Newton approach, in which analytical first derivatives and approximate second derivatives are used to search for a lower energy point on the PES. In this method, a Newton–Raphson step, $\Delta\mathbf{x}$, is taken on a local quadratic PES:

$$\Delta\mathbf{x} = -\mathbf{H}^{-1}\mathbf{g} \quad (1)$$

where \mathbf{g} is the gradient (first derivative) and \mathbf{H} is the Hessian (second derivative). In practical implementations, the Newton–Raphson step is stabilized with a control technique such as the rational function optimization (RFO)^{3,4} and the trust radius model (TRM).^{5–8} In the quasi-Newton approach for geometry optimizations, computationally expensive ana-

lytical evaluations of the second derivatives are replaced with a numerical Hessian update scheme, such as BFGS,^{9–12} SR1,¹³ and PSB.^{14,15} To obtain an optimization step $\Delta\mathbf{x}$, eq 1 can be solved with a direct inversion of the Hessian or RFO/TRM in the eigenvector space of the Hessian. However, inversion or diagonalization of a Hessian matrix incurs an $O(N^3)$ scaling, where N is the number of nuclear degrees of freedom. Such a cubic scaling can become a substantial bottleneck in the optimization of large molecules with semiempirical or linear scaling electronic structure methods. Alternatively, an iterative $O(N^2)$ approach can be carried out to search for the RFO solution in the reciprocal space of the Hessian.¹⁶ However, iterative solutions are often associated with numerical instabilities and a large scaling prefactor. In addition, iterative solutions do not offer direct computation of eigenvectors and eigenvalues that are important in vibrational analysis and transition state optimizations.

On the other hand, the choice of coordinate system in which the geometry optimization is conducted is crucial for successful convergence of the geometry optimization algorithm. Generally, geometry optimization in an appropriate set of internal coordinates can converge significantly faster than in Cartesian coordinates.^{17–19} However, a practical internal coordinate system for molecular geometry optimiza-

* Corresponding author e-mail: li@chem.washington.edu.

[†] University of Washington, Seattle.

[‡] Kunming University of Science and Technology.

[§] Gaussian Inc.

tions usually contains redundancy. For large-scale systems, coordinate redundancy can become the speed-limiting factor arising from operating on excessively large matrices. In principle, redundancy can be removed by transformation to the nonredundant internal coordinate, leading to a much smaller coordinate space and less computationally expensive matrix inversion. However, obtaining the redundant–nonredundant vectors is another $O(N^3)$ procedure where N is the number of nonredundant coordinates. This dilemma largely prevents a practical application of geometry optimization in nonredundant coordinate space.

In this paper, we present an eigenspace update scheme with an $O(N^2)$ scaling in the nonredundant internal coordinate space. Computational performance and efficiency are compared for a select set of large molecules with the standard full diagonalization-based Bery algorithm^{18,20} in the redundant internal coordinate with RFO.

II. Methodology

A. Eigenspace Update—An $O(N^2)$ Algorithm for Molecular Geometry Optimization. Assume eigenvectors, \mathbf{C} , and eigenvalues, λ , of the Hessian exist at step i :

$$\mathbf{H}_i = \mathbf{C}_i \cdot \lambda_i \cdot \mathbf{C}_i^\dagger \quad (2)$$

Then, a forward optimization step, $\Delta \mathbf{x}_i$, can be obtained by means of RFO or TRM, using eq 1, resulting in a new geometry, \mathbf{x}_{i+1} ; a new gradient, \mathbf{g}_{i+1} ; and a new and updated Hessian, \mathbf{H}_{i+1} . In the current implementation, we use a weighted combination of BFGS^{9–12} and SR1¹³ with the square root of the Bofill²¹ weighting factor (see refs 22 and 23 for details). The new Hessian \mathbf{H}_{i+1} can be projected into the previous eigenvector space \mathbf{C}_i as

$$\Delta_{i+1} = \mathbf{C}_i^\dagger \cdot \mathbf{H}_{i+1} \cdot \mathbf{C}_i \quad (3)$$

Equation 3 can be considered as an intermediate diagonalization step. In principle, one can obtain the eigenvalues, λ_{i+1} , and eigenvectors, \mathbf{C}_{i+1} , of the Hessian by diagonalizing Δ_{i+1} :

$$\lambda_{i+1} = \mathbf{A}_{i+1}^\dagger \cdot \Delta_{i+1} \cdot \mathbf{A}_{i+1} \quad (4)$$

$$\mathbf{C}_{i+1} = \mathbf{C}_i \cdot \mathbf{A}_{i+1} \quad (5)$$

However, eqs 4 and 5 do not initially seem advantageous over the traditional approach of direct diagonalization of \mathbf{H}_{i+1} . From the molecular vibration standpoint, nonzero off-diagonal elements in eq 3 are related to vibrational couplings and anharmonicities. For any given normal mode, there exists a vibration which gives rise to the strongest coupling, or the largest off-diagonal element in Δ_{i+1} . Usually, in a nearly quadratic potential energy surface, changes in the Hessian matrix are small. If we only consider changes of the Hessian from the strongest couplings, Δ_{i+1} in eqs 3 and 4 can be replaced with a tridiagonal form, $\Delta_{3,i+1}$, where the only nonzero off-diagonal elements are the first diagonal below/above the main diagonal. Diagonalization of a tridiagonal matrix in eq 4 formally scales as $O(N^2)$ when the *Divide and Conquer*²⁴ algorithm is employed. As a result, eqs 3–5

become an eigenvector and eigenvalue update scheme, which is much more efficient than direct diagonalization.

In the following tests, we use a LAPACK subroutine to obtain eigenvalues and eigenvectors of a tridiagonal matrix $\Delta_{3,i+1}$. The projected Hessian matrix Δ_{i+1} is reorganized by swapping rows/columns in every optimization step so that the largest off-diagonal element for any given mode is positioned in the first diagonal below/above the main diagonal. The reorganization starts from the first projection vector in \mathbf{C}_i . When the projected Hessian matrix is reorganized, the related projection eigenvectors \mathbf{C}_i are also rearranged consistently according to the rows/columns being swapped in Δ_{i+1} . Note that this reorganization does not change the map between eigenvectors \mathbf{C} and the geometric coordinates \mathbf{x} .

B. Transformation to Nonredundant Coordinate Space. Analytical gradients, \mathbf{g} , are usually computed in the Cartesian coordinate and require geometries, \mathbf{x} , represented in the same coordinate as well. The transformation from the Cartesian coordinate \mathbf{x} to the redundant internal coordinate \mathbf{q} can be done with a symmetric \mathbf{G} matrix built from the Wilson \mathbf{B} matrix:²⁵

$$\mathbf{B} = \frac{d\mathbf{q}}{d\mathbf{x}} \quad (6)$$

$$\mathbf{G} = \mathbf{B}\mathbf{B}^T \quad (7)$$

where $d\mathbf{q}$ and $d\mathbf{x}$ are infinitesimal displacements in internal and Cartesian coordinates, respectively. With the transformation matrices defined in eqs 6 and 7, the gradient and optimization step can be transformed between two representations (Cartesian and redundant internal):

$$\mathbf{f}_q = \mathbf{G}^{-1}\mathbf{B}\mathbf{f}_x \quad (8)$$

$$\Delta \mathbf{x} = \mathbf{B}^T \mathbf{G}^{-1} \Delta \mathbf{q} \quad (9)$$

where \mathbf{f}_x and \mathbf{f}_q are forces in the Cartesian and redundant internal coordinate, and the Newton–Raphson step, $\Delta \mathbf{q}$, in the redundant internal coordinate is

$$\Delta \mathbf{q} = \mathbf{H}^{-1}\mathbf{f} \quad (10)$$

Note, in the quasi-Newton approach, the Hessian matrix can be updated in redundant internal coordinates without transformation back to the Cartesian coordinate.

For optimizations of large-scale systems, a smaller nonredundant coordinate space is preferred. The redundancy condition can be determined by single value decomposition (SVD) of the matrix \mathbf{G} :¹⁹

$$\mathbf{G} = (\mathbf{K}\mathbf{L}) \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} (\mathbf{K}\mathbf{L})^T, \Lambda \neq 0 \quad (11)$$

In eq 11, \mathbf{K} corresponds to the nonredundant coordinate space with nonzero eigenvalues Λ , and \mathbf{L} consists of redundant eigenvectors of \mathbf{G} . However, obtaining the eigenspace of the \mathbf{G} matrix is another $O(N^3)$ procedure and speed-limiting step, and large molecules often have a large number of redundant coordinates. The eigenspace update concept introduced in section II.A can be used here to reduce the

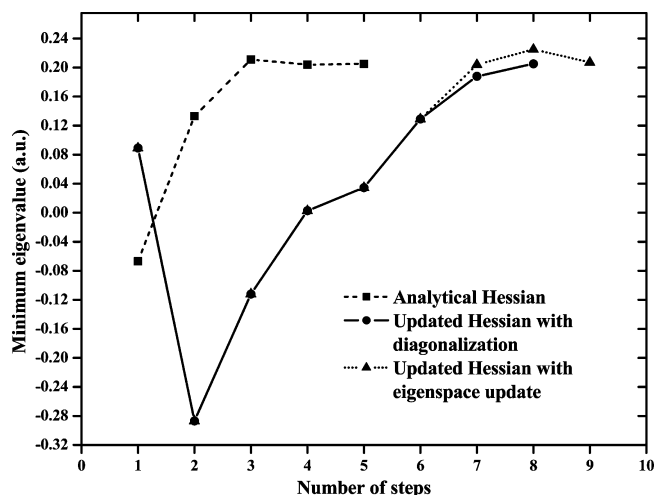


Figure 1. Comparison of minimum eigenvalues of the Hessian during optimization of a single water molecule, using analytical Hessian, updated Hessian with diagonalization, and Hessian eigenspace update approaches.

dimension of the \mathbf{G} matrix. Assuming the linear dependence in the internal coordinate space remains the same from step to step, the full \mathbf{G} matrix can then be reduced into the nonredundant coordinate space using the previous nonredundant vectors \mathbf{K} , followed by an SVD of the reduced and much smaller $\tilde{\mathbf{G}}$ matrix.

$$\tilde{\mathbf{G}} = \mathbf{K}_i^T \cdot \mathbf{G}_{i+1} \cdot \mathbf{K}_i \quad (12)$$

$$\tilde{\mathbf{G}} = \mathbf{U}_{i+1} \cdot \Lambda_{i+1} \cdot \mathbf{U}_{i+1}^T \quad (13)$$

The nonredundant coordinate space at the $i + 1$ step can be constructed accordingly:

$$\mathbf{K}_{i+1} = \mathbf{K}_i \cdot \mathbf{U}_{i+1} \quad (14)$$

The SVD of the full \mathbf{G} matrix in redundant internal space is performed only once in the first step. Subsequent geometry optimization steps will take advantage of this eigenspace update scheme (eqs 12–14) to obtain the nonredundant coordinate space \mathbf{K}_{i+1} . As the reduced $\tilde{\mathbf{G}}$ matrix is significantly smaller than the original \mathbf{G} matrix in redundant space, the SVD on $\tilde{\mathbf{G}}$ is no longer computationally dominant. As a result, obtaining the redundant–nonredundant transformation matrix \mathbf{K}_{i+1} becomes a pseudo- $O(N^2)$ approach.

Table 1. Comparison of Computational Costs for RFO Approach Using the Diagonalization in Redundant Space and Eigenspace Update Method^a

molecules (numbers of atoms)	energy (au)	diagonalization in redundant space			eigenspace update		
		geom ^b	SCF	steps ^c	geom ^b	SCF	steps ^c
hydrazobenzene (26)	0.129468	1.00	8.70	20	0.52	8.70	20
taxol (113)	-0.666862	1.00	0.68	63	0.45	0.71	66
for-(Ala) ₁₀ -NH ₂ (106)	-0.733344	1.00	0.82	100	0.66	0.97	115
for-(Ala) ₂₀ -NH ₂ (206)	-1.424445	1.00	0.63	206	0.52	0.70	238
for-(Ala) ₂₅ -NH ₂ (259)	-1.779332	1.00	0.48	82	0.38	0.50	87
crambin (642)	-4.167923	1.00	0.27	389			
crambin (642)	-4.169380				0.20	0.23	333

^a The computational cost is evaluated using the total CPU time of geometry steps using full diagonalization in redundant internal coordinate as the unit reference. ^b One geometry step includes forming the Wilson \mathbf{B} matrix; obtaining the nonredundant eigenspace, Hessian update, diagonalization, or eigenspace update; and solving the RFO equation. ^c Total number of geometry optimization steps.

With the eigenspace of the nonredundant internal coordinate, the Newton–Raphson step in eq 10 can be transformed into the nonredundant internal coordinate space with the following equation:

$$\mathbf{K}^{-1} \Delta \mathbf{q} = (\mathbf{K}^T \mathbf{H} \mathbf{K})^{-1} \mathbf{K}^T \mathbf{f} \quad (15)$$

If we define the Newton–Raphson step, Hessian, and force in the nonredundant internal coordinate as

$$\Delta \tilde{\mathbf{q}} = \mathbf{K}^{-1} \Delta \mathbf{q} \quad (16)$$

$$\tilde{\mathbf{H}} = \mathbf{K}^T \mathbf{H} \mathbf{K} \quad (17)$$

$$\tilde{\mathbf{f}} = \mathbf{K}^T \mathbf{f} \quad (18)$$

eq 15 becomes the familiar form of the Newton–Raphson step, but in the nonredundant internal coordinate space:

$$\Delta \tilde{\mathbf{q}} = \tilde{\mathbf{H}}^{-1} \tilde{\mathbf{f}} \quad (19)$$

The RFO correction can be applied to eq 19 with the Hessian eigenspace update scheme presented in the previous section. The displacement is then transformed back to redundant internal coordinate:

$$\Delta \mathbf{q} = \mathbf{K} \Delta \tilde{\mathbf{q}} \quad (20)$$

followed by another transformation to the Cartesian coordinate using the curvilinear eq 9 through an iterative approach.¹⁹

III. Benchmarks and Discussion

Optimizations are carried out using the AM1 Hamiltonian as implemented in the development version of the Gaussian program²⁶ with the addition of the geometry optimization algorithm using the eigenspace update (ESU) method in the nonredundant internal coordinate presented in sections II.A and B. For all test cases, the geometry optimization is considered converged when the maximum component of the force vector is less than 4.5×10^{-4} au, the root-mean-square (RMS) force is less than 3×10^{-4} au, the maximum component of the geometry displacement is less than 1.8×10^{-3} au, and the RMS geometry displacement is less than 1.2×10^{-3} au. To ensure a smooth convergence, the tridiagonal approximation of the reduced Hessian matrix is turned on when the regular RFO correction is smaller than one-tenth of the minimum

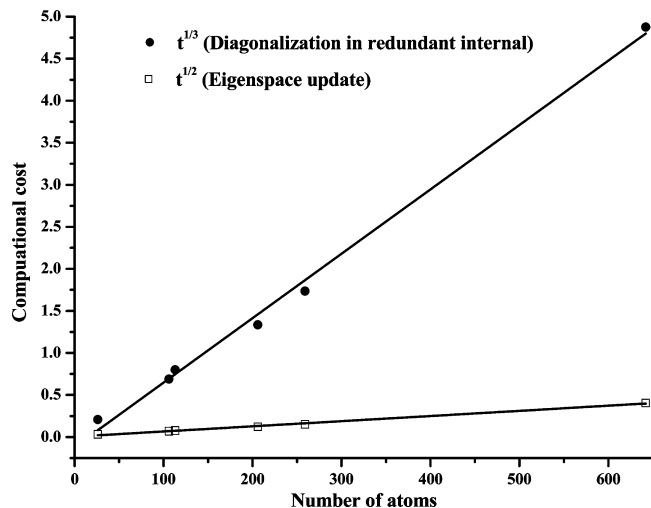


Figure 2. Comparison of computational costs using diagonalization in the redundant coordinate and eigenspace update. The computational cost of a single optimization step is plotted against the number of atoms.

Hessian eigenvalue. In the following discussion, we refer to a *geometry step* as the procedure including (1) forming the Wilson **B** matrix; (2) obtaining the nonredundant eigenspace (section II.B), (3) Hessian update, and (4) diagonalization or eigenspace update (section II.A); and (5) solving the RFO equation. For semiempirical and force field methods, the computational cost of the analytical gradient is not considered to be a computationally expensive step.

In the ESU approach, the Hessian eigenspace is usually an approximation and requires a number of optimization steps to converge to the true value. Figure 1 shows the convergence of the Hessian eigenspace using the ESU method and direct diagonalization of the updated Hessian, compared to the true analytical Hessian. It is known that the Hessian update scheme is able to converge to the true Hessian within a few geometry steps. Built on the Hessian update scheme, the Hessian eigenspace update method (section II.A) adds an additional degree of approximation. Therefore, the convergence behavior of the ESU approach is slower than the diagonalization-based method, but only by a few geometry steps. Nevertheless, the gain in computational speed owing to the $O(N^2)$ scaling and the nonredundancy is promising for large scale systems.

Table 1 lists relative computational costs for geometry optimizations using the ESU method for a select set of molecules compared to those obtained with full diagonalization in the redundant internal coordinate. The computational cost is evaluated using the total CPU time of geometry steps of the Hessian diagonalization-based RFO approach as the unit reference. For smaller molecules, such as hydrazobenzene, there are no savings in the overall computational cost. Although the computational cost for geometry steps is reduced, the approximate nature of the Hessian eigenspace in ESU leads to several additional optimization steps compared to the diagonalization-based RFO approach. As a result, additional computational cost incurs, arising from additional SCF

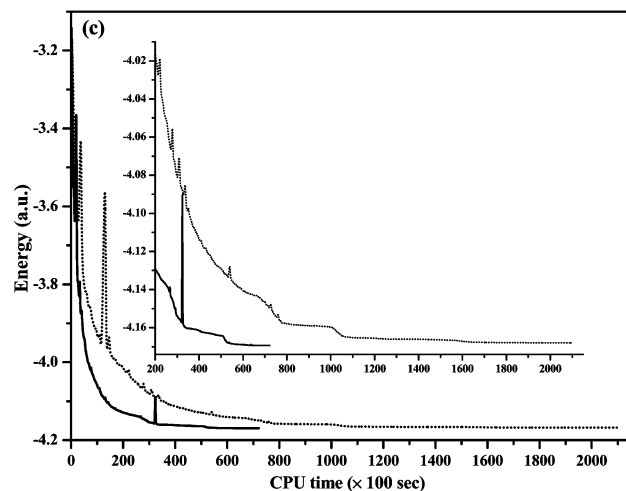
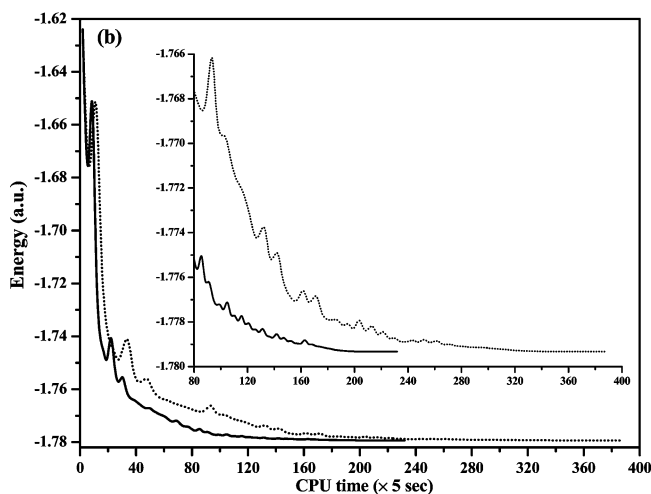
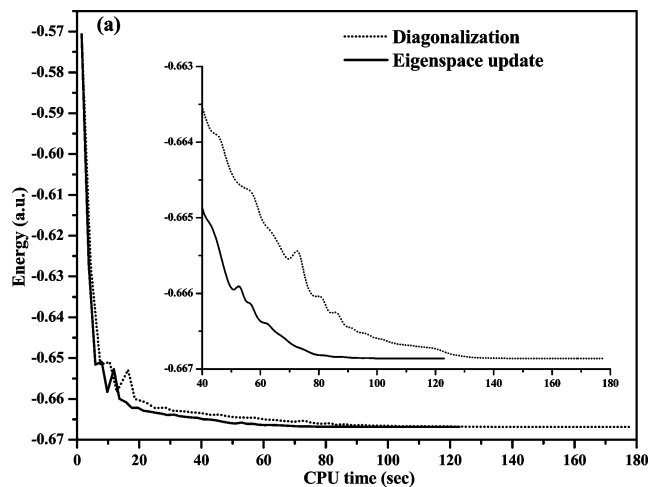


Figure 3. Comparison of optimizations using diagonalization in redundant space and eigenspace update methods for (a) taxol, (b) 25-alanine chain, and (c) crambin at the AM1 level of theory. See Table 1 regarding the evaluation of computational cost.

steps, and therefore the application of ESU for small molecules is not particularly advantageous.

On the other hand, as the molecular size increases, the cost for geometry steps becomes dominant in computational cost for semiempirical self-consistent field (SCF) or force

field energy calculations. In Table 1, we include test cases of large-scale molecules, such as taxol, alanine chains, and crambin. Because of excessive nuclear and electronic degrees of freedom, and numerous undesirable shallow potential wells on the PES, these molecules are often difficult and computationally expensive to optimize. As the nuclear degrees of freedom increase, the computational cost of the ESU-based method is noticeably less expensive than the conventional approach using full diagonalization in redundant coordinates. For the 25-alanine chain case, a ~60% computational saving is observed. In this case, the total computational cost for SCF iterations in the ESU-based method is slightly (~1%) more expensive than the conventional approach due to a slightly larger number of geometry steps. Therefore, such a large computational savings in the ESU-based method can be ascribed to the efficient eigenspace update algorithm in nonredundant internal coordinate space introduced herein. Although we cannot make a direct comparison for the largest test case, crambin, as the two methods converge to different minima,²⁷ a factor of 3 in computational cost is definitely noticeable.

To further understand the computational performance of the ESU approach, we plot in Figure 2 the computational cost of a single geometry step as a function of the number of atoms. It is clear that ESU is an $O(N^2)$ method while the diagonalization-based approach exhibits an $O(N^3)$ scaling. As the molecular sizes increase, the advantage of using an $O(N^2)$ approach becomes highly appreciated. Figure 3a–c illustrate optimization processes of selected large-sized molecules: taxol, alanine-25, and crambin at the AM1 level of theory. It shows that the ESU method takes a similar optimization pathway as diagonalization in the redundant-space-based RFO approach but has the advantage of being much more efficient.

IV. Conclusion

This paper presents a geometry optimization method using an eigenspace update approach. This method takes advantage of previously computed eigenvectors for obtaining eigenspaces of the transformation \mathbf{G} matrix and the Hessian matrix and exhibits an $O(N^2)$ scaling. This method shows an encouraging efficiency for geometry optimization, with up to a factor of 3 savings in computational cost for large-sized molecular systems. The optimization pathways are similar to those using conventional diagonalization in the redundant-space-based RFO approach. An even more promising implementation of the ESU method would be combination direct inversion in the iterative subspace algorithm (DIIS), as exemplified by the energy-represented DIIS²³ and the simultaneous DIIS methods.²⁷

Acknowledgment. This work was supported by the U.S. National Science Foundation (CHE-CAREER 0844999 to X.L.). Additional support from Gaussian Inc. and the University of Washington Student Technology Fund is gratefully acknowledged.

References

- Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514.
- Hratchian, H. P.; Schlegel, H. B. Finding Minima, Transition States, and Following Reaction Pathways on Ab Initio Potential Energy Surfaces. In *Theory and Applications of Computational Chemistry: The First 40 Years*; Dykstra, C. E., Kim, K. S., Frenking, G., Scuseria, G. E., Eds.; Elsevier: New York, 2005; p 195.
- Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52.
- Simons, J.; Nichols, J. *Int. J. Quantum Chem.* **1990**, *24*, 263.
- Murray, W.; Wright, M. H. *Practical Optimization*; Academic: New York, 1981.
- Powell, M. J. D. *Non-linear Optimization*; Academic: New York, 1982.
- Dennis, J. E.; Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Non-linear Equations*; Prentice Hall: New York, 1983.
- Scales, L. E. *Introduction to Non-linear Optimization*; Macmillan: Basingstoke, U. K., 1985.
- Broyden, C. G. *J. Inst. Math. Appl.* **1970**, *6*, 76.
- Fletcher, R. *Comput. J. (Switzerland)* **1970**, *13*, 317.
- Goldfarb, D. *Math. Comput.* **1970**, *24*, 23.
- Shanno, D. F. *Math. Comput.* **1970**, *24*, 647.
- Murtagh, B.; Sargent, R. W. H. *Comput. J. (Switzerland)* **1972**, *13*, 185.
- Powell, M. J. D. *Nonlinear Programming*; Academic: New York, 1970.
- Powell, M. J. D. *Math. Program.* **1971**, *1*, 26.
- Farkas, O.; Schlegel, H. B. *J. Chem. Phys.* **1999**, *111*, 10806.
- Baker, J. J. *Comput. Chem.* **1993**, *14*, 1085.
- Peng, C. Y.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **1996**, *17*, 49.
- Pulay, P.; Fogarasi, G. *J. Chem. Phys.* **1992**, *96*, 2856.
- Schlegel, H. B.; Yarkony, D. R. Geometry optimization on potential energy surfaces. In *Modern Electronic Structure Theory*; World Scientific: Singapore, 1995; p 459.
- Bofill, J. M. *J. Comput. Chem.* **1994**, *15*, 1.
- Farkas, O.; Schlegel, H. B. *Phys. Chem. Chem. Phys.* **2002**, *4*, 11.
- Li, X.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 835.
- Trefethen, L. N.; Bau, D. *Numerical Linear Algebra*; Society for Industrial and Applied Mathematics: Philadelphia, PA, 1997.
- Wilson, E. B. *Molecular Vibrations; the Theory of Infrared and Raman Vibrational Spectra*; McGraw-Hill: New York, 1955.
- Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X. H.; Hratchian, P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi,

J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Parandekar, P. V.;

Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian Development*, version H.01; Gaussian, Inc.: Wallingford, CT, 2009.

(27) Moss, C. L.; Li, X. *J. Chem. Phys.* **2008**, *129*, 114102.

CT100214X

JCTC

Journal of Chemical Theory and Computation

Significant van der Waals Effects in Transition Metal Complexes

Per E. M. Siegbahn,^{*,†} Margareta R. A. Blomberg,[†] and Shi-Lu Chen^{†,‡}

Department of Physics, ALBA NOVA and Department of Biochemistry and Biophysics, Arrhenius Laboratory, Stockholm University, SE-106 91, Stockholm, Sweden, and School of Science, Beijing Institute of Technology, Beijing 100081, P.R. China

Received April 22, 2010

Abstract: There is, in general, very good experience using hybrid DFT to study mechanisms of enzyme reactions containing transition metals. For redox reactions, the B3LYP* functional, which has 15% exact exchange, has been shown to be particularly accurate. Still, there are some cases which have turned out to be quite difficult with large errors. In the present study, the effects of van der Waals interaction have been investigated for these cases, using the empirical formula of Grimme. The results are encouraging.

Introduction

Hybrid density functional theory has been an extremely successful tool in studying mechanisms for enzymatic reactions involving transition metals.^{1–3} Barriers and reaction energies within 3–5 kcal/mol from experimental results have generally been found, provided the chemical model is large enough. Still, there are continuous reports of failures of hybrid DFT for transition metal complexes. For example, the energy differences between the peroxo and bis- μ -oxo isomers of copper dimer complexes appear to show big errors of 10–15 kcal/mol.⁴ Also, the binding of methyl and adenosyl to cobalamin⁵ as well as the binding of small molecules to heme groups have been reported to be underestimated by the same, or even larger, magnitude.⁶ The most common explanation for the DFT failures has been the inability to describe multireference effects, since DFT is inherently a single determinantal method.^{6,7} That explanation, implying that B3LYP should very often be distrusted for transition metal complexes, is in sharp contrast to the excellent experience obtained when studying chemical reactions with this method. In the present letter, these failures for transition metal complexes are reinvestigated using recent improvements of hybrid DFT, where van der Waals effects are empirically included.⁸ A significant advantage of this improvement is that it can be applied on top of the results

of a well established DFT method, such as B3LYP.⁹ Since it has been argued that the fraction of exact exchange is a way to tune nondynamical correlation effects in DFT,¹⁰ another advantage is that these effects and van der Waals effects, which have different origins, can be separated. This type of improvement is in contrast to suggestions to improve the results in difficult cases by selecting a different functional depending on the problem investigated. In that approach, the hybrid B3LYP functional could be used for molecules containing first and second row atoms, while a nonhybrid functional like BP86 should be used for binding methyl and adenosyl to cobalamin, and M06 functionals should be used for copper complexes.^{4–7} There are some previous investigations on the inclusion of explicit van der Waals effects for 3d transition metal complexes. For example, significant improvements were demonstrated for noncovalent ligand binding energies in some chromium complexes.¹¹ In a benchmark test containing 3d transition metals, the inclusion of a dispersion correction on top of the B97 functional essentially removed cases with large errors.¹² Also, the M06-L functional has been demonstrated to show much improved results for noncovalent interactions in 3d complexes.¹³

Methods

In the calculations performed here, the B3LYP* functional¹⁴ has been used if not otherwise indicated. This is a slight modification of the original B3LYP functional⁹ with 15% exact exchange (rather than 20%), which has been found to

* Corresponding author e-mail: ps@physto.se.

[†] Stockholm University.

[‡] Beijing Institute of Technology.

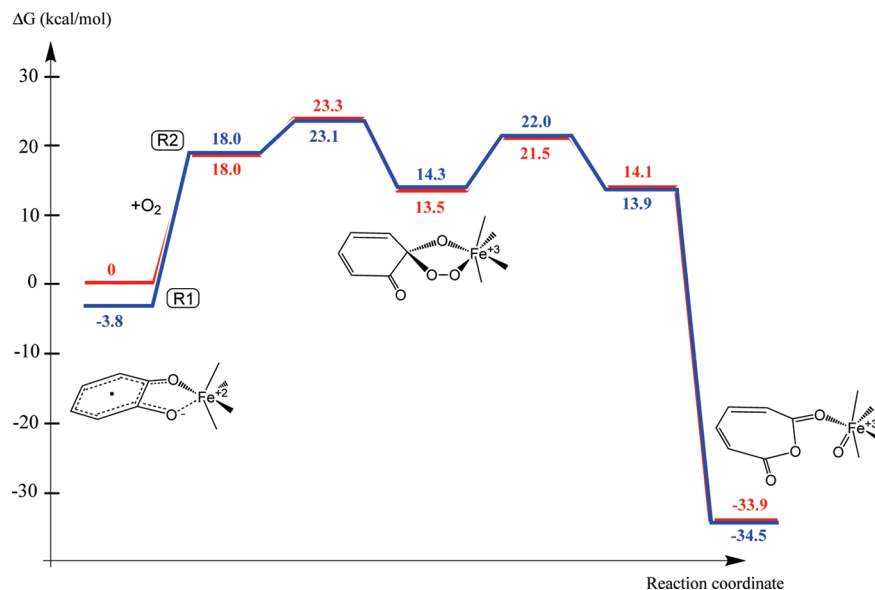


Figure 1. Energy diagram for the biomimetic intradiol-cleaving dioxygenase discussed in the text. The red line is without, the blue line with van der Waals effects.

be superior in most cases for describing oxidations of transition metals.^{2,14,15} The energetics discussed were obtained using large, nearly saturated basis sets (cc-pvtz(-f)) in single point calculations at geometries optimized using a smaller basis set. Solvent effects were included with a dielectric constant chosen from case to case. They were not found to be significant in the reactions discussed below. The calculations were performed using the Jaguar program.¹⁶

A Typical Energy Diagram. The discussion of the results will start with a typical example of a reaction mechanism involving a transition metal complex, taken from a recent application.¹⁷ This reaction is for a biomimetic intradiol-cleaving dioxygenase, where the details do not matter in the present context. The energy diagram obtained at the B3LYP* level for the suggested mechanism is shown in Figure 1. The starting point of the reaction is an Fe(III) complex with a bound catechol substrate. In the first step of the mechanism in the figure, O_2 binds to the complex. The mechanism then proceeds by formation of a bridging peroxide, and the cleavage of the O–O bond. Finally, the ring of the catechol substrate is opened in between the carbons carrying the hydroxyl groups, an intradiol cleavage. The competing mechanism is an extradiol ring opening, and the main question asked is why one mechanism is preferred and not the other. This and other questions were answered by the model calculations, and the barriers computed were reasonable compared to experiments. In short, there was no sign of any failure of B3LYP* in spite of the redox chemistry occurring.

Also shown in Figure 1 are the relative effects (set to zero for **R2**) from adding van der Waals interactions through the empirical formula. The most striking feature is that these effects are almost perfectly constant throughout the reaction, except for the step from **R1** to **R2** when O_2 becomes bound, where there is a significant relative effect of -3.8 kcal/mol, which is expected since two additional atoms are added to the complex. The conclusion drawn is that, apart from this effect, which has been pointed out before,^{18,19} the energy

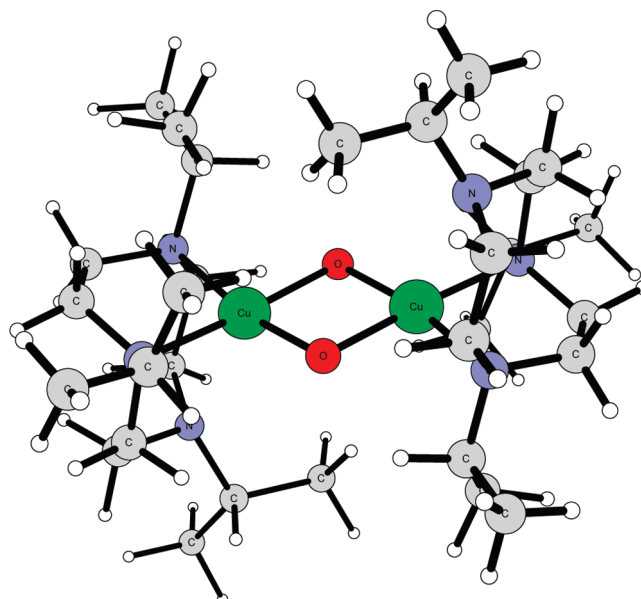


Figure 2. Optimized $Cu_2(III,III)$ -bis- μ -oxo structure with iPr_3TACN ligands.

diagram is almost unaffected by adding van der Waals effects. The most important mechanistic issues, such as the choice of the intradiol or extradiol pathway, are therefore also unaffected.

Dicopper Complexes. The discussion of dicopper complexes will start with the case with iPr_3TACN ligands, shown in Figure 2. This is an interesting system experimentally, for which it has been shown that the bis- μ -oxo (in the figure) and the peroxo complexes are in equilibrium.²⁰ The experimental estimate of the energy difference is 0.9 kcal/mol favoring the peroxo complex. The previously calculated B3LYP value is 15.0 kcal/mol, favoring the peroxo complex, and was claimed to show “some of the worst agreement between pure and hybrid functionals” ever reported.⁴ The pure functional value was in good agreement with experiments, and this type of functional was therefore strongly

recommended for these systems. The present B3LYP* value is 4.2 kcal/mol. Adding the van der Waals effect of -3.6 kcal/mol leads to an energy difference of only 0.6 kcal/mol. The solvent effects (already included) favor the bis- μ -oxo structure by -1.7 kcal/mol. The zero-point effects of $+2.0$ kcal/mol (not included) favor the peroxo structure, while the relativistic effects of -3.9 kcal/mol (not included)²¹ are in the opposite direction. Overall, the present result is in quite reasonable agreement with experiments. The difference between the present and the previous results is partly explained by the use of B3LYP*, which favors the bis- μ -oxo structure by 5.7 kcal/mol compared to B3LYP, and the van der Waals effects of 3.6 kcal/mol, also favoring the bis- μ -oxo structure. However, this is not the full explanation, since these corrections only sum up to 9.3 kcal/mol and the difference amounts to 14.4 kcal/mol. The larger basis set used here also appears to play a role. For the corresponding case with $i\text{Pr}_2\text{TACD}$ ligands, the present calculations favor the peroxo structure by 5.3 kcal/mol, including a van der Waals effect of -5.8 kcal/mol. This result is also in qualitative agreement with experiments, even though the exact energy difference is not known in that case. For another type of ligand, termed DBED, the experimental energy difference is shifted slightly (by 1 kcal/mol) toward the peroxo complex, compared to the case with $i\text{Pr}_3\text{TACN}$ ligands.²² The present calculations give an energy difference of 2.7 kcal/mol favoring the peroxo structure, where the van der Waals contribution is -2.8 kcal/mol. The calculated difference from the case with $i\text{Pr}_3\text{TACN}$ ligands is thus $+2.1$ kcal/mol in favor of the peroxo structure, in comparison with the experimental preference by about $+1.0$ kcal/mol.

Since the above results are in such good agreement with experiments, it is interesting to investigate what the same level of treatment gives for the energy difference between the peroxo and the bis- μ -oxo structure of the dicopper complexes appearing in the hemocyanin, tyrosinase, and catechol oxidase enzymes, which has been a strongly debated issue.²³ In these cases, there are three histidine ligands on each copper. Modeling these by imidazoles leads to an energy difference of 14.1 kcal/mol favoring the peroxo structure, which includes a van der Waals effect of -3.2 kcal/mol. This result is in qualitative agreement with the fact that only the peroxo complex has been observed. The previous conclusion that the bis- μ -oxo structure does not enter into the mechanisms in these enzymes,²³ therefore, still appears to hold.

Binding of Methyl and Adenosyl to Cobalamin. The cleavage of the Co–C bond in methyl- or adenosyl-cobalamin, see Figure 3, is a common first step in many reactions catalyzed by enzymes including the vitamin B₁₂ cofactor. In the case of adenosyl, the cleavage is homolytic, and the resulting radical abstracts a hydrogen atom from the substrate in the second step. In the case of methyl, the cleavage is heterolytic, resulting in a methyl cation. A major problem in quantum chemical studies of these enzymes has been that it has turned out to be difficult to obtain a proper homolytic bond dissociation energy for the Co–C bond. B3LYP has been found to underestimate the bond strength by 10–15 kcal/mol. In contrast, nonhybrid methods like

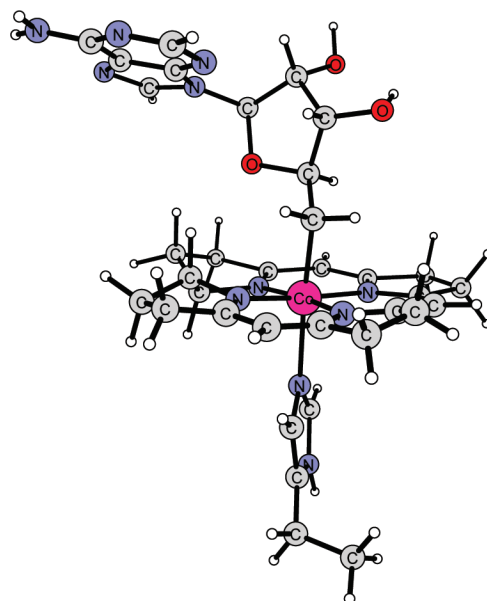


Figure 3. Optimized structure for cobalamin with a bound adenosyl ligand.

BP86 have given values much closer to experiments but have instead had problems in describing the energetics of the subsequent reaction steps.

Using the model in Figure 3, a B3LYP value for the Co–C bond strength for methyl of 16.2 kcal/mol (including zero-point and solvent corrections) was obtained in ethylene-glycol ($\epsilon = 40$),²⁴ compared to the experimental value of 37 ± 3 kcal/mol.²⁵ A discrepancy of as much as 20 kcal/mol is thus obtained, in line with previous bad experience using hybrid DFT. At the B3LYP* level, the bond strength increases to 20.7 kcal/mol, which is still a severe underestimation. However, in this case, the van der Waals effects turn out to be quite large with 11.3 kcal/mol increasing the bond strength. The resulting bond strength of 32.0 kcal/mol is at least in reasonably good agreement with experiments. For adenosyl, the corresponding results are 16.7 kcal/mol at the B3LYP* level, and 29.5 kcal/mol with van der Waals effects added. This result agrees very well with the experimental value of 30 kcal/mol.²⁶ The van der Waals contribution of 12.8 kcal/mol for adenosyl is remarkably large. The reason is the large number of rather short atom–atom distances between the substrate and the cobalamin. The van der Waals contribution from the metal is very small due to the cutoff value in the empirical formula. It is clear that the mechanisms of these cobalamin-containing enzymes cannot be described without van der Waals interactions.

Small Molecule–Heme Interactions. The binding of molecular oxygen to heme-iron is important in several biological processes, for example, in oxygen transport and in respiration. Also, the binding of other small molecules, such as CO and NO, to heme-iron plays an important role, for example, as inhibitors for O₂ binding. As has been pointed out before, the binding energies of these small molecules to heme-Fe(II) as calculated using the B3LYP functional are significantly too small,⁶ at the same time as it was shown that CASPT2 calculations gave good agreement with experimental results. Since this failure of the DFT calculations

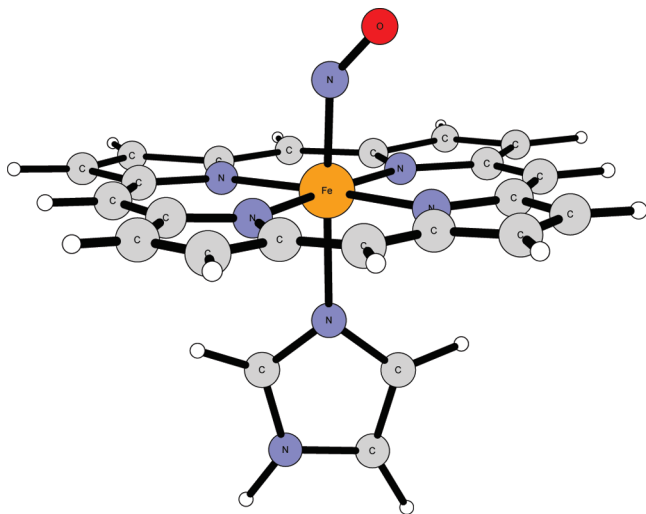


Figure 4. Model used in the calculations of small molecule–heme complexes.

Table 1. Calculated Fe–X Binding Energies in Six-Coordinate Heme, Where X is CO, NO, or O₂ (Zero-Point Effects Included)

	Fe–CO (kcal/mol)	Fe–NO (kcal/mol)	Fe–O ₂ (kcal/mol)
B3LYP	10.9	7.6	1.1 (5.9) ^c
B3LYP-D	20.6	16.9	8.8 (13.6) ^c
B3LYP*	16.6	16.3	5.5 (10.3) ^c
B3LYP*-D	26.3	25.6	13.2 (18.0) ^c
exp ^a	19.5 (18.5)	22.8 (22.8)	10.1 (16.1)
exp ^b	18.1		12.3

^aFrom dissociation barriers in myoglobin corrected for “the protein effect”,⁶ uncorrected myoglobin values in parentheses. ^bDissociation barriers in protoheme. ^cValues within parentheses are spin-corrected.

can be due to the lack of both multireference and van der Waals effects, new calculations have been performed using both the B3LYP and the B3LYP* functionals, and adding the empirical van der Waals corrections according to Grimme, giving rise to four calculated binding energies for each system: B3LYP, B3LYP-D, B3LYP*, and B3LYP*-D. The model used in the calculations is shown for the NO case in Figure 4, and the results are summarized in Table 1.

As can be seen from Table 1, the attractive van der Waals effects are quite significant, 9.7 kcal/mol for CO, 9.3 kcal/mol for NO, and 7.7 kcal/mol for O₂. This is due to the interaction between the binding molecule and the large number of atoms in the heme group. The B3LYP-D values therefore come quite close to the experimental values for both CO and O₂, while the binding of NO is still much too small, 16.9 kcal/mol compared to the experimental value of 22.8 kcal/mol. Reducing the amount of exact exchange in going from B3LYP to B3LYP* gives quite large increases in the binding energy, 5.7 kcal/mol for CO, 8.6 kcal/mol for NO, and 4.4 kcal/mol for O₂. In the case of (heme)Fe–O₂, which has an antiferromagnetically coupled open shell singlet ground state, there is also a calculated spin-correction of 4.8 kcal/mol, and in the table all values are given with and without this spin correction. The spin correction and the reduction of exact exchange are partly related effects, connected to the multiconfigurational character of the wave function, and if both effects are applied, the calculated

binding energy is clearly too large, 18.0 kcal/mol (with van der Waals correction) as compared to the experimental values of 10–12 kcal/mol. Without spin correction, the B3LYP*-D value of 13.2 kcal/mol is in good agreement with the experimental values. For NO, the B3LYP*-D value of 25.6 kcal/mol is a bit too large compared to the experimental value (22.8 kcal/mol), but it is still the best calculated value for NO binding.

In summary, for these heme systems, the situation is more complicated than for the other systems discussed above. The van der Waals effects are large and important, but for some systems, B3LYP-D gives better agreement with experimental values, while for others, B3LYP*-D gives better results. Further investigations are therefore needed to find out how these systems should be best described.

Conclusions

In a few important cases, taken from studies of enzyme mechanisms, van der Waals effects have been shown to be quite significant. In most cases they appear in the step where a substrate becomes bound to the metal cofactor. Apart from this step, the van der Waals effects are normally small. By including the van der Waals effects, and reducing the amount of exact exchange to 15%, the results are in good agreement with experiments even for most of these difficult cases. The most significant improvements appear for the binding of adenosyl to cobalamin and for biomimetic dicopper complexes, where very good results are obtained. Interestingly, the previous exclusion of the Cu₂(III,III) state in the mechanism of tyrosinase²³ still appears to hold. The exceptions to the excellent results are the cases when a small molecule is bound to a heme group, where the results are still not quite satisfactory. For these systems, more work and experience are needed to improve the situation.

Supporting Information Available: Coordinates for all the structures discussed in the present paper. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Siegbahn, P. E. M.; Borowski, T. *Acc. Chem. Res.* **2006**, *39*, 729–738.
- (2) Siegbahn, P. E. M. *J. Biol. Inorg. Chem.* **2006**, *11*, 695–701.
- (3) Siegbahn, P. E. M.; Himio, F. *J. Biol. Inorg. Chem.* **2009**, *14*, 643–651.
- (4) Lewin, J. L.; Heppner, D. E.; Cramer, C. J. *J. Biol. Inorg. Chem.* **2007**, *12*, 1221–1234. Gherman, B. F.; Cramer, C. J. *Coord. Chem. Rev.* **2009**, *253*, 723–753.
- (5) Jensen, K. P.; Ryde, U. *J. Phys. Chem. A* **2003**, *107*, 7539–7545.
- (6) Radon, M.; Pierloot, K. *J. Phys. Chem. A* **2008**, *112*, 11824–11832.
- (7) Cramer, C. J.; Wloch, M.; Piecuch, P.; Puzzarini, C.; Gagliardi, L. *J. Phys. Chem. A* **2006**, *110*, 1991–2004.
- (8) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108. Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.

- (9) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (10) Friesner, R. A.; Knoll, E. H.; Cao, Y. *J. Chem. Phys.* **2006**, *125*, 124107.
- (11) Minenkov, Y.; Occhipinti, G.; Jensen, V. R. *J. Phys. Chem. A* **2009**, *113*, 11833–11844.
- (12) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (13) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.
- (14) Reiher, M.; Salomon, O.; Hess, B. A. *Theor. Chem. Acc.* **2001**, *107*, 48–55.
- (15) Siegbahn, P. E. M. *Chem.—Eur. J.* **2008**, *27*, 8290–8302.
- (16) *Jaguar 5.5*; Schrödinger, LLC: Portland, OR, 1991–2003.
- (17) Georgiev, V.; Noack, H.; Borowski, T.; Blomberg, M. R. A.; Siegbahn, P. E. M. *J. Phys. Chem. B* In press.
- (18) Wirstam, M.; Lippard, S. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2003**, *125*, 3980–3987.
- (19) Lundberg, M.; Morokuma, K. *J. Phys. Chem. B* **2007**, *111*, 9380–9389.
- (20) Cahoy, J.; Holland, P. L.; Tolman, W. B. *Inorg. Chem.* **1999**, *38*, 2161–2168.
- (21) Flock, M.; Pierloot, K. *J. Phys. Chem. A* **1999**, *103*, 95–102.
- (22) Mirica, L. M.; Vance, M.; Rudd, D. J.; Hedman, B.; Hodgson, K. O.; Solomon, E. I.; Stack, T. D. P. *Science* **2005**, *308*, 1890–1892.
- (23) Siegbahn, P. E. M.; Wirstam, M. *J. Am. Chem. Soc.* **2001**, *123*, 11819–11820.
- (24) Chen, S.-L.; Siegbahn, P. E. M.; Blomberg, M. R. A. *Biochemistry* submitted.
- (25) Martin, B. D.; Finke, R. G. *J. Am. Chem. Soc.* **1990**, *112*, 2419. Martin, B. D.; Finke, R. G. *J. Am. Chem. Soc.* **1992**, *114*, 585.
- (26) Finke, R. G.; Hay, B. P. *Inorg. Chem.* **1984**, *23*, 3041–3043. Hay, B. P.; Finke, R. G. *J. Am. Chem. Soc.* **1986**, *108*, 4820–4829. Garr, C. D.; Finke, R. G. *Inorg. Chem.* **1993**, *32*, 4414–4421.

CT100213E

JCTC

Journal of Chemical Theory and Computation

Equipartition and the Calculation of Temperature in Biomolecular Simulations

Michael P. Eastwood, Kate A. Stafford,[§] Ross A. Lippert, Morten Ø. Jensen, Paul Maragakis, Cristian Predescu, Ron O. Dror, and David E. Shaw^{*†}

D. E. Shaw Research, New York, New York 10036

Received June 5, 2009

Abstract: Since the behavior of biomolecules can be sensitive to temperature, the ability to accurately calculate and control the temperature in molecular dynamics (MD) simulations is important. Standard analysis of equilibrium MD simulations—even constant-energy simulations with negligible long-term energy drift—often yields different calculated temperatures for different motions, however, in apparent violation of the statistical mechanical principle of equipartition of energy. Although such analysis provides a valuable warning that other simulation artifacts may exist, it leaves the actual value of the temperature uncertain. We observe that Tolman's generalized equipartition theorem should hold for long stable simulations performed using velocity-Verlet or other symplectic integrators, because the simulated trajectory is thought to sample almost exactly from a continuous trajectory generated by a shadow Hamiltonian. From this we conclude that all motions should share a single simulation temperature, and we provide a new temperature estimator that we test numerically in simulations of a diatomic fluid and of a solvated protein. Apparent temperature variations between different motions observed using standard estimators do indeed disappear when using the new estimator. We use our estimator to better understand how thermostats and barostats can exacerbate integration errors. In particular, we find that with large (albeit widely used) time steps, the common practice of using two thermostats to remedy so-called hot solvent–cold solute problems can have the counter-intuitive effect of causing temperature imbalances. Our results, moreover, highlight the utility of multiple-time step integrators for accurate and efficient simulation.

1. Introduction

Fueled by algorithmic improvements and by the growth of computer power, molecular dynamics (MD) simulations are making increasingly important scientific contributions to biology. There is considerable interest in further accelerating simulations and improving their accuracy. Since most biomolecular simulations are of classical systems at equilibrium, one useful measure of accuracy is the extent to which

the distribution of energy among different degrees of freedom is consistent with the equipartition theorem of statistical mechanics.¹ The most familiar consequence of equipartition is that each particle in an equilibrium system has an average kinetic energy of $k_B T/2$ (where T is the temperature of the system and k_B is the Boltzmann constant) arising from its motion in each spatial dimension. This result does not depend on details of the potential energy function (the “force field”). In biomolecular simulations, thermalization of kinetic energy typically occurs on a subnanosecond time scale,^{2,3} so substantial deviations from equipartition in long simulations are likely symptoms of a problem with the simulation methodology. One symptom, whose presence is often tested for in practice, is a difference between the temperature of the solvent and solute, often referred to as a *hot solvent–cold solute* problem.⁴ A hot solvent–cold solute problem could

^{*} Corresponding author. E-mail: David.Shaw@DEShawResearch.com.

[†] David E. Shaw is also affiliated with the Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032.

[§] Current address: Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032.

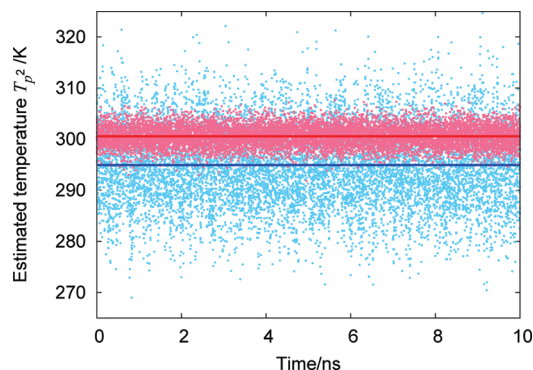


Figure 1. A hot solvent–cold protein problem. Temperatures of the protein ubiquitin (blue points) and water solvent (red points) are shown as a function of simulation time. Data were taken from an all-atom constant energy simulation that used velocity-Verlet integration with a 2 fs time step and bonds to hydrogen constrained; more details are given in Section 3.2. Temperatures were evaluated using the squares of the on-step velocity-Verlet momenta. The lines denote the average temperature values over the entire simulation.

have many potential causes, and testing for the presence of this symptom (and similar deviations from equipartition) has helped to diagnose underlying problems in barostats,⁵ thermostats^{6,7} and in approximate treatments of long-range electrostatic^{8–10} and dispersive¹¹ interactions and has led to various methodological improvements.

Here, we investigate how truncation errors arising from the finite simulation time step δ_t affect equipartition and the calculation of temperature. We mainly focus on the widely used velocity-Verlet integrator,^{12,13} but the basic theoretical finding applies to symplectic integrators in general. To help rule out the well-documented causes of a breakdown of equipartition noted above,^{5–11} we initially focus on constant-energy simulations that are stable (that is, those that show little long-term energy drift). Fortunately, this is achievable with currently typical simulation parameters, so the results are directly relevant in practice. Although not as dramatic as hot solvent–cold solute problems arising from other origins, the effects of truncation errors can still be substantial. In Figure 1, for example, the temperature of a protein and the surrounding water molecules are shown as a function of simulation time. The temperature of the protein is seen to be lower than that of the water by about 6 K.

Such results are widely understood to expose real simulation artifacts originating in the finite integration time step, but it is unclear whether they reflect any actual temperature differences or even whether temperature has a precise definition for $\delta_t > 0$. In this paper, we show how the definition of temperature generalizes to $\delta_t > 0$ and show that, in examples like the one above, different motions do share a single temperature. Our reasoning is straightforward. The velocity-Verlet integrator is symplectic¹⁴ and is, thus, thought to sample positions and momenta almost exactly from a trajectory generated by a modified Hamiltonian,¹⁵ often called a *shadow Hamiltonian*. Since particle momenta do not enter this Hamiltonian quadratically, the equipartition relation is not applicable. We expect, however, that generalized equipartition,^{1,16} which holds for a broad class of

Hamiltonians, will be applicable to the shadow Hamiltonian. This implies the existence of a single well-defined simulation temperature for all motions that is given by the product of momentum and velocity and that is easily evaluated in practice.

To avoid potential confusion at the outset and to clarify why estimating temperature from the product of momentum p and velocity v (a “ pv formula”) is distinct from previous approaches, we emphasize the finding¹⁷ that when $\delta_t > 0$, $v \neq p/m$, where m is mass; we also explain more carefully what we mean here by velocity and momentum. By momentum, we mean the canonical momentum that enters the shadow Hamiltonian. This is directly provided by the integrator. For velocity Verlet, the momentum is simply the usual on-step velocity-Verlet momentum. By velocity, we mean the *instantaneous* rate of change of position on the underlying trajectory generated by the shadow Hamiltonian. The integrator directly yields positions (and momenta) but not their time derivatives, and the velocity cannot be exactly expressed in terms of a finite number of positions and momenta. In particular, as noted above, $v \neq p/m$, even though p/m might commonly be called a velocity; for velocity Verlet, p/m (often called the on-step “velocity-Verlet velocity”) differs from the velocity by $\mathcal{O}(\delta_t^2)$. Thus if only the on-step velocity-Verlet momenta are used to evaluate temperature (a “ p^2 formula,” as used in Figure 1), then temperatures will also be in error by order δ_t^2 . Nevertheless, it is straightforward to construct more accurate velocity estimators.¹⁸ One simple approach is a polynomial interpolation over positions sampled at different times; the velocity estimator appearing in Beeman’s version of Verlet¹³ is a well-known special case, and unsurprisingly it is possible to increase the accuracy further by interpolating over more positions. Perhaps counterintuitively, however, our pv formula shows that having obtained an accurate estimate of v , the temperature follows via the product pv (even though p/m may itself be a poor estimator of v) and *not* the square of the accurately estimated velocity (a “ v^2 formula”), as appears to be typically assumed.^{18,19} Indeed, as will become clearer below, if temperature is estimated using highly accurate velocities alone, a hot solvent–cold protein problem, like that shown in Figure 1, will simply be replaced by a cold solvent–hot protein problem of similar magnitude.

We test our theoretical conclusions numerically for two systems. First, we calculate the temperatures of vibrational and translational motion in a diatomic fluid as a function of δ_t . We find that p^2 and v^2 temperature estimates each yield substantially different values for the two motions, but the pv estimator shows that the temperatures of these motions are in fact identical within a very small statistical error. Thus although conventional equipartition (by which we mean the usual, as opposed to generalized, equipartition relation) breaks down, generalized equipartition holds, and a well-defined temperature exists. Using analytical estimates, we confirm that although deviations from conventional equipartition do not reflect temperature differences, they do reflect the real difference between the Hamiltonian and its shadow. Second, we perform all-atom MD simulations of ubiquitin in explicit solvent, examining the temperature of the different quasiharmonic protein motions and comparing the overall

temperature of the protein to the solvent. Again, we find different motions to share a single temperature, even when conventional equipartition is not satisfied.

In addition to providing an accurate estimator for simulation temperature and using it to confirm that generalized equipartition is satisfied in stable simulations, we use it to investigate how integration errors can be exacerbated by use of a thermostat or barostat. This danger has been recently highlighted;¹⁸ we discuss it in light of the new estimator and demonstrate some potential pitfalls. Notably, we use our estimator to show that unless the time step is chosen to be sufficiently small, use of multiple thermostats can lead to a breakdown of generalized equipartition with genuine temperature imbalances and with heat flow in the system. The commonly used remedy of hot solvent–cold solute problems in which one thermostat is applied to protein and another to the solvent, in an effort to maintain them at the same temperature, can thus potentially have a counterintuitive, and counterproductive, effect.

The root cause of all the simulation artifacts investigated in this paper is truncation error. Satisfying generalized equipartition by no means implies that the simulation is free from this source of artifacts; indeed, the breakdown of conventional equipartition signals their existence. Reducing the time step naturally reduces truncation error and brings the different temperature estimators into agreement, but due to the computational expense of MD, this solution is often unpalatable. One promising approach to reduce errors is to modify the integration scheme. Using the deviations from conventional equipartition as a criterion, we show, for example, that for our test systems, the reversible reference system propagation algorithm (r-RESPA) multiple-time step scheme²⁰ can achieve the benefit of a reduced velocity-Verlet time step at a fraction of the computational expense.

2. Theory

To generalize the definition of simulation temperature to $\delta_t > 0$, we make use of two established concepts: generalized equipartition and the shadow Hamiltonian, which we briefly review in Sections 2.1 and 2.2, respectively. In Section 2.3, we give our definition of simulation temperature for $\delta_t > 0$. Using the harmonic oscillator as an analytically tractable example, we quantify errors in some conventional estimates of simulation temperature in Section 2.4. Appendix A describes how to estimate temperatures of motions that involve multiple atoms, such as quasiharmonic motions in proteins. The effect of integration errors on simulation pressure is discussed in Appendix B.

2.1. Generalized Equipartition. Generalized equipartition (eqs 2 and 3) was derived by Tolman,¹⁶ who considered the following canonical ensemble average for the Hamiltonian system $H(\mathbf{p}, \mathbf{q})$:

$$\begin{aligned} \left\langle x_i \frac{\partial H}{\partial x_i} \right\rangle &= \frac{1}{Q} \int dx \left(\frac{-x_i}{\beta} \right) \frac{\partial e^{-\beta H}}{\partial x_i} \\ &= k_B T - \frac{1}{\beta Q} \int dx_1, \dots, dx_{i-1} dx_{i+1}, \dots, dx_{2N} [x_i e^{-\beta H}]_{x_{\min}}^{x_{\max}} \end{aligned} \quad (1)$$

We use x_i to label an element of either position or momentum. N_f is the number of positional degrees of

freedom, $\beta = 1/k_B T$, and $Q = \int dx \exp(-\beta H)$. Under the relatively mild requirement that the surface (second) term on the right-hand side vanishes, use of Hamilton's equations leads to the exact result:

$$\langle p_i \dot{q}_i \rangle = k_B T \quad (2)$$

$$-\langle q_i \dot{p}_i \rangle = k_B T \quad (3)$$

where q_i and p_i label individual positions and conjugate momenta respectively, and the dot denotes a time derivative. The velocities are

$$v_i \equiv \dot{q}_i = \frac{\partial H}{\partial p_i} \quad (4)$$

We have presented Tolman's original proof for the canonical (NVT) ensemble, because of its brevity. For the microcanonical (NVE) ensemble, which is relevant to our development below, the proof is described elsewhere;¹ the result is identical, apart from an i -independent correction of order N_f^{-1} .²¹ The use of periodic boundary conditions means linear momentum is often conserved in simulations. This constraint leads to a modification to eq 2 of order N^{-1} , where N is the number of particles in the simulation.^{21,22} We ignore effects of this magnitude except where explicitly noted. More importantly, if q_i is a periodic coordinate, then the second term on the right-hand side of eq 1 may be nonzero, in which case eq 3 will not hold. A commonly encountered example is for simulations using periodic boundary conditions in which the position q_i of an atom is restricted to values that lie within the simulation box; assuming that the Hamiltonian is translationally invariant, then knowledge of a single positional coordinate q_i provides no information about \dot{p}_i , so evidently these quantities are uncorrelated, and $\langle q_i \dot{p}_i \rangle = \langle q_i \rangle \langle \dot{p}_i \rangle = 0$, as may also be demonstrated by explicitly evaluating the second term on the right-hand side of eq 1.

In MD simulations we usually use a Hamiltonian of the form

$$H_0(\mathbf{p}, \mathbf{q}) = U(\mathbf{q}) + \frac{1}{2} \mathbf{p}^T \mathbf{m}^{-1} \mathbf{p} \quad (5)$$

where \mathbf{q} denotes atom positions, \mathbf{m} is the diagonal mass matrix ($\mathbf{m}_{ij} = \delta_{ij} m_i$), and U is the force field. Since H_0 contains only quadratic terms in p_i , eq 4 shows that—for exact trajectories—velocities and momenta are related through $m_i v_i = p_i$. Eq 2 thus reduces to the familiar form of kinetic energy equipartition:

$$\langle m_i v_i^2 \rangle_0 = k_B T \quad (6)$$

$$\langle p_i^2 / m_i \rangle_0 = k_B T \quad (7)$$

where the 0 is used to emphasize that a Hamiltonian of the form H_0 is assumed. The shadow Hamiltonian corresponding to H_0 need not take this form, however, as we review below.

MD simulations are often performed subject to holonomic constraints, for example, to keep certain bond lengths fixed. Since eq 2 assumes an unconstrained ensemble average, it does not directly apply. In principle, one can construct a new Hamiltonian describing the dynamics of a system subject to

N_c constraints by finding $3N - N_c$ unconstrained generalized positions and their conjugate momenta; eq 2 will then apply. For our purposes, however, it is sufficient to establish two results. First, for any subset A of position coordinates that are not involved in a constraint with position coordinates outside that subset, we find—after some algebra—that a result similar to eq 2 holds

$$\sum_{i \in I_A} \langle p_i v_i \rangle = (N_{f:A} - N_{c:A}) k_B T \quad (8)$$

Here I_A contains the indices of the coordinates in A, and $N_{f:A} - N_{c:A}$ is the number of positional degrees of freedom in A minus the number of constraints to which they are subject. Second, in the common case where the constraints are functions only of interatomic distance, it is straightforward to identify some unconstrained generalized coordinates. Formally, a subset B of the positions in A may be identified and transformed to any linear average coordinate $Q_B = \sum_{i \in I_B} w_i q_i$, with $\sum_{i \in I_B} w_i = 1$ and $N_{f:B} - 1$ relative coordinates, such that Q_B is unconstrained. Thus

$$\langle P_B V_B \rangle = k_B T \quad (9)$$

where $V_B = \dot{Q}_B$ and $P_B = \sum_{i \in I_B} p_i$ is the momentum conjugate to Q_B . Typically, Q_B will be a center-of-mass coordinate. Although eqs 8 and 9 are intuitively obvious, we mention them here to make explicit that, like eq 2, they hold without need to assume a Hamiltonian of the form H_0 . Only when a Hamiltonian of form H_0 is assumed can they be written in terms of squared velocities or momenta, as in eqs 6 and 7.

2.2. Shadow Hamiltonian. We focus on the velocity-Verlet integration of H_0 , for which

$$\mathbf{q}(t + \delta_t) = \mathbf{q}(t) + \delta_t \mathbf{m}^{-1} \mathbf{p}(t) + \frac{\delta_t^2}{2} \mathbf{m}^{-1} \mathbf{F}(t) \quad (10)$$

$$\mathbf{p}(t + \delta_t) = \mathbf{p}(t) + \frac{\delta_t}{2} (\mathbf{F}(t) + \mathbf{F}(t + \delta_t)) \quad (11)$$

where the elements of the force vector are $F_i = -\partial U / \partial q_i$, as usual. This integration scheme is symplectic, a property that can be maintained in the presence of holonomic constraints.²³ A symplectic integrator is one for which the mapping $(\mathbf{p}(t), \mathbf{q}(t)) \rightarrow (\mathbf{p}(t + \delta_t), \mathbf{q}(t + \delta_t))$ is a canonical transformation, just as it is for continuous Hamiltonian dynamics. This suggests there might be a Hamiltonian whose *exact* dynamics generates the flow $(\mathbf{p}(t), \mathbf{q}(t)) \rightarrow (\mathbf{p}(t + \delta_t), \mathbf{q}(t + \delta_t))$. This shadow Hamiltonian H_δ is expected to be similar but not equal to H_0 , whose *approximate* dynamics generates the same flow. Finding H_δ is a problem of backward error analysis.^{15,24} One may construct an asymptotic expansion for H_δ by adding terms to H_0 to create a Hamiltonian whose exact dynamics matches that of eqs 10 and 11 order by order in δ_t . Since velocity Verlet is symmetric (reversing the sign of the time step gives the inverse method), only even powers of δ_t appear

$$H_\delta(\mathbf{p}, \mathbf{q}) = H_0(\mathbf{p}, \mathbf{q}) + \frac{\delta_t^2}{2!} \delta H^{(2)}(\mathbf{p}, \mathbf{q}) + \frac{\delta_t^4}{4!} \delta H^{(4)}(\mathbf{p}, \mathbf{q}) + \dots \quad (12)$$

It is possible to construct accurate numerical approximations for the correction terms;²⁵ we note that an estimator (eq 68 of ref 26) for $\delta H^{(2)}$ with errors of order δ_t^2 (and thus an estimator of H_δ with errors of order δ_t^4) has existed for many years in the CHARMM code,²⁷ where it is called a “high-frequency correction.” Except for particularly simple forms of H_0 , there is no guarantee eq 12 converges, but numerical tests to high orders have found the conservation of the shadow Hamiltonian to improve when successively higher order terms are included.²⁵

The second-order term in eq 12 is¹⁷

$$\delta H^{(2)} = \frac{1}{6} (\mathbf{m}^{-1} \mathbf{p})^T \mathbf{K} (\mathbf{m}^{-1} \mathbf{p}) - \frac{1}{12} \mathbf{F}^T \mathbf{m}^{-1} \mathbf{F} \quad (13)$$

where the Hessian has elements $K_{ij} = \partial^2 U / \partial q_i \partial q_j$. Since \mathbf{K} depends on \mathbf{q} , even at second-order, H_δ has a different form than H_0 , and the velocities and momenta are thus no longer related through a simple mass factor.¹⁷ Specifically, eq 4 gives

$$\mathbf{v} = \mathbf{m}^{-1} \mathbf{p} + \frac{\delta_t^2}{6} \mathbf{m}^{-1} \mathbf{K} \mathbf{m}^{-1} \mathbf{p} + \mathcal{O}(\delta_t^4) \quad (14)$$

2.3. Equipartition for the Shadow Hamiltonian. The nontrivial relationship between velocities and momenta (eq 14) for the shadow Hamiltonian has immediate consequences for equipartition. In particular, eq 2 clearly no longer precisely reduces to eqs 6 and 7. This suggests that the simulation temperature should be defined using $T \equiv T_{pv}$, where

$$k_B T_{pv} = \langle p_i v_i \rangle_{\delta_t} \quad (15)$$

and the δ_t subscript emphasizes that the ensemble average depends on the time step through the shadow Hamiltonian. With this definition, the temperature for all motions $\{i\}$ will be the same, if generalized equipartition is satisfied. The alternative quantities

$$k_B T_{p^2} = \langle p_i^2 / m_i \rangle_{\delta_t} \quad (16)$$

$$k_B T_{v^2} = \langle m_i v_i^2 \rangle_{\delta_t} \quad (17)$$

differ from T by an amount $\mathcal{O}(\delta_t^2)$. We use the terms pv , p^2 , or v^2 formula to refer to any method of calculating the temperature that is in the spirit of eqs 15, 16, or 17, respectively. (The p^2 formula corresponds to the usual method of obtaining temperature when using velocity-Verlet integration, that is using only on-step momenta provided by the integrator.) As demonstrated below, it is also possible to calculate the temperature, making use of eq 3, provided appropriate (nonperiodic) coordinates are used. Note that the temperature, as defined in eq 15, is a property of the distribution sampled during the simulation, rather than a property of the desired distribution that would have been sampled in the small-time-step limit. Statistical reweighting,^{24,28} while a powerful tool to infer the desired ensemble from the one sampled, thus cannot be directly applied to sampled p^2 values to obtain $T \equiv T_{pv}$. The temperature one would obtain by such a reweighting is instead the (known)

temperature of the desired ensemble. This has been explicitly demonstrated (for NVT simulations using a Nosé–Poincaré thermostat) in numerical experiments where reweighting T_{pv^2} accurately recovered the thermostat’s target temperature.²⁴

To calculate the temperature from eq 15, one needs to calculate both velocities and momenta. The momenta \mathbf{p} are, by construction, directly available at every step of the simulation. The velocities $\mathbf{v} \equiv \dot{\mathbf{q}} \neq \mathbf{m}^{-1}\mathbf{p}$ are not but can be accurately estimated from several consecutive positions via interpolation, as described in Section 3.3. Since the standard terminology we have adopted can lead to confusion, we emphasize that velocity-Verlet samples *momenta*—specifically, the canonical momenta of the shadow Hamiltonian—and not velocities. Naturally, this state of affairs is unaltered if eqs 10 and 11 are explicitly written in terms of what are called velocity-Verlet velocities, $\mathbf{v}_{vv} \equiv \mathbf{m}^{-1}\mathbf{p}$. As its definition shows, \mathbf{v}_{vv} is simply and precisely related to the *momenta* but differs from the velocities $\mathbf{v} \equiv \dot{\mathbf{q}}$, when $\delta_t > 0$.

In Section 2.1, we made a few comments on the applicability of the generalized equipartition formula. Here we make some related comments on the applicability of eq 15 to MD simulations. First, eq 15 was derived assuming that a shadow Hamiltonian exists, i.e., that the dynamics generated by the integrator is the exact dynamics of some underlying Hamiltonian. This is strictly true for certain simple forms of H_0 but not for biomolecular force fields, where the asymptotic expansion eq 12 does not converge. For simulations with small energy drift, however, we find much encouragement in earlier work²⁵ that Hamiltonians defined by truncating eq 12 can describe the dynamics generated by the integrator extremely accurately. Nevertheless, an important part of this paper is to test numerically whether generalized equipartition holds. Second, in addition to the usual statistical error, estimators for the temperature, based on eq 15, contain errors from the velocity interpolation. As discussed further below, n^{th} -order polynomial interpolation essentially leads to $\mathcal{O}(\delta_t^n)$ errors, thus sufficiently high-order interpolation can make the errors in the estimated temperature negligible for practical purposes. Third, thermostats and barostats used in MD simulations may entail modification to the equations of motion such that they are no longer of Hamiltonian form (this is the case for both Nosé–Hoover²⁹ and Berendsen³⁰ thermostats). Although a rigorous analysis of such effects on eq 15 appears possible for some thermostats, in this paper we make the simplifying assumption that any modifications can be ignored. This is intuitively reasonable for thermostats coupled to a large number of degrees of freedom and is borne out by our numerical results on ubiquitin, which are very similar for NVE and NVT simulations. Similarly, although we focus on straightforward MD simulations here, we expect eq 15 to be relevant to Monte Carlo sampling methods that use molecular dynamics, such as parallel tempering.³¹ Finally, we note that, while eq 15 was derived assuming a symplectic integrator, milder conditions are sufficient. In particular, if the integrator conserves phase-space volume and has a conserved quantity $\tilde{H}_\delta(\mathbf{p}, \mathbf{q})$, which need not be a Hamiltonian, then eq 15 will hold if $\partial\tilde{H}_\delta/\partial p_i = v_i$.

2.4. Harmonic Oscillator. We briefly illustrate the above results for the simple case of a one-dimensional harmonic oscillator with mass m and spring constant k :

$$H_0^{\text{osc}}(p, q) = \frac{1}{2}kq^2 + \frac{p^2}{2m} \quad (18)$$

As is well-known (for example, see ref 17), the shadow Hamiltonian is a harmonic oscillator with modified mass and spring constant,

$$H_{\delta_t}^{\text{osc}}(p, q) = \frac{1}{2}k_\delta q^2 + \frac{p^2}{2m_\delta} \quad (19)$$

The modified parameters m_δ and k_δ are given by

$$\frac{m_\delta}{m} = \frac{\omega}{\omega_{\delta_t}} \sqrt{1 - \left(\frac{1}{2}\omega\delta_t\right)^2} \approx 1 - \frac{1}{6}(\omega\delta_t)^2 \quad (20)$$

$$\frac{k_\delta}{k} = \frac{\omega_{\delta_t}}{\omega} \sqrt{1 - \left(\frac{1}{2}\omega\delta_t\right)^2} \approx 1 - \frac{1}{12}(\omega\delta_t)^2 \quad (21)$$

where the modified frequency $\omega_{\delta_t} = (k_\delta/m_\delta)^{1/2}$ is

$$\frac{\omega_{\delta_t}}{\omega} = \frac{\arcsin\left(\frac{1}{2}\omega\delta_t\right)}{\frac{1}{2}\omega\delta_t} \approx 1 + \frac{1}{24}(\omega\delta_t)^2 \quad (22)$$

Both the shadow mass and spring constant decrease with increasing time step, and vanish as $\omega\delta_t \rightarrow 2$, which coincides with the stability limit of the integrator.

Suppose that the harmonic oscillator is weakly coupled to a heat bath at temperature T . The quantities T_{pv} , T_{p^2} , and T_{v^2} defined in the previous subsection are related to T as follows

$$\frac{T_{pv}}{T} = 1; \quad \frac{T_{p^2}}{T} = \frac{m_\delta}{m}; \quad \frac{T_{v^2}}{T} = \frac{m}{m_\delta} \quad (23)$$

Whereas T_{pv} correctly shows the oscillator to have the same temperature as the bath, T_{p^2} underestimates and T_{v^2} overestimates the temperature by the same factor. The disagreement between the three estimators provides a valuable indication of the magnitude of truncation error. In addition, we may also use eq 3; if we define $k_B T_{qp} = -\langle q\dot{p} \rangle_\delta$, and $k_B T_{qF} = \langle q(kq) \rangle_\delta$, then

$$\frac{T_{qp}}{T} = 1; \quad \frac{T_{qF}}{T} = \frac{k}{k_\delta} \quad (24)$$

Finally, since some codes have estimators of the shadow energy available, it is of interest to estimate the temperature obtained using the difference of the shadow energy and the potential energy. With the definition $k_B T_{H-U}/2 = \langle H_\delta \rangle - \langle kq^2/2 \rangle$, we find

$$\frac{T_{H-U}}{T} = 2 - \frac{k}{k_\delta} \quad (25)$$

showing that the difference of the shadow energy and potential energy yields a temperature estimator with errors of order δ_t^2 .

3. Simulation and Analysis Details

All simulations were performed using Desmond.³² Periodic boundary conditions were used, and center-of-mass motion was removed every time step. Every picosecond, coordinates, and velocities were saved for nine consecutive time steps to allow interpolation. Since we wish to isolate truncation errors from round-off errors, Desmond and analysis programs used double-precision arithmetic. Energy drift was small: even assuming it is converted entirely into kinetic energy, the drift over the entire length of each *NVE* simulation corresponds to a temperature change of less than 0.2 K. All errorbars denote statistical errors, which were estimated using a blocking method.³³

3.1. Diatomic Fluid. The simulated system contained 1000 oxygen-like diatomic molecules of 32 atomic mass units in a cubic box with a side length of 44.42 Å. The force field contained only two kinds of terms, each a simple function of interatomic distance r : intramolecular bond stretch terms of the form $k(r - r_0)^2/2$, and van der Waals interactions of Lennard-Jones form $4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6]$ between atoms on different molecules; there were no electrostatic interactions. Most parameters ($r_0 = 1.15$ Å, $\sigma = 2.9$ Å, and $\epsilon = 0.586152$ kJ/mol) were taken to be the OPLS-AA/L force field³⁴ values for oxygen. The bond force constant was reduced by approximately a factor of six from the OPLS-AA/L value for oxygen to $k = 1368$ kJ/mol/Å², which corresponds to a lengthened bond-vibration period of 48.05 fs. These parameters ensure that energy transfer between the high-frequency vibrations and the other motions is rapid (equipartition is reached in hundreds of ps). The van der Waals interactions were truncated at 10 Å and calculated using a neighbor list of pairs within 11.25 Å that was updated every ~15 fs. We performed 10 *NVE* simulations of 400 ns using the velocity-Verlet integration scheme with time steps ranging from 0.5 to 5 fs. We also performed 9 *NVE* simulations of 400 ns using the r-RESPA integrator²⁰ with bond stretches calculated every 0.5 fs and with intermolecular interactions between 2 and 10 times less frequently. All simulations started from the same configuration (which had been pre-equilibrated using a thermostat at 300 K) and were each assigned different initial velocities that were randomly chosen from a Maxwell–Boltzmann distribution.

3.2. Ubiquitin. The set up and parameters for the ubiquitin simulations were similar to those used previously.³⁵ PDB entry 1D3Z³⁶ was solvated with 5302 explicit water molecules, giving a total of 17 137 atoms in a cubic box of side 55.71 Å. We used the OPLS-AA/L all-atom force field,³⁴ as implemented in GROMACS version 3.1.4,³⁷ for the protein, together with the SPC water model.³⁸ Electrostatic forces were computed using the particle mesh Ewald method³⁹ with a screening Gaussian width of $10/(3\sqrt{2}) \approx 2.36$ Å and with fifth-order interpolation to a cubic mesh of $64 \times 64 \times 64$ points; real-space contributions to the electrostatics and van der Waals interactions were truncated at 10 Å and calculated from a list of pairs separated by less than 11 Å that was assembled every ~12 fs. Water molecules and lengths of bonds to hydrogens were rigidly constrained using M-SHAKE,⁴⁰ as implemented⁴¹ in Desmond. Energy

minimization and equilibration under conditions of constant temperature and pressure yielded a conformation, from which 8 simulations of 11 ns were started. These simulations were as follows: three *NVE* velocity-Verlet simulations with different time steps (1.25, 2, and 2.5 fs); one *NVE* simulation using an r-RESPA multiple time step scheme,²⁰ in which nonbonded interactions were evaluated every 2.5 fs and the remaining interactions every 1.25 fs; four simulations using a velocity-Verlet time step of 2 fs and coupled in different ways to Berendsen thermostats³⁰ with relaxation times of 0.5 ps (either a single thermostat coupled to the entire system, or the protein or water alone, or two independent thermostats with the first coupled to the water and the second to the protein).

To calculate the temperature of the protein, T_{pv}^{protein} , we made use of eq 8, where i ran over all protein atoms and Cartesian dimensions. The temperatures of the water and the entire system, T_{pv}^{water} and T_{pv}^{system} , were calculated analogously. The quantities $T_{p^2}^{\text{water}}$, $T_{p^2}^{\text{protein}}$, $T_{p^2}^{\text{system}}$ were calculated in the same way, except that the interpolated velocity was replaced by the corresponding momenta divided by the mass; this is how temperatures of components are normally calculated in simulation. We also calculated the analogous v^2 quantities. The position coordinates used for the quasiharmonic analysis were center-of-mass coordinates of protein heavy atoms and their covalently bonded hydrogen atoms. This gives a total of 1800 position coordinates when overall rotation and translation are excluded.

3.3. Interpolation. While positions and momenta are directly available from the integrator at every time step, their time derivatives are not, but can be estimated by interpolation. Here we use straightforward polynomial interpolation. For velocities, fitting an n^{th} -order polynomial through the positions at the time of interest and the following $n/2$ and preceding $n/2$ times and taking the time derivative gives a time-symmetric n^{th} -order approximation to the velocities, $\mathbf{v}^{(n)}$. The second-order result

$$\mathbf{v}^{(2)}(t) = (\mathbf{x}(t + \delta_t) - \mathbf{x}(t - \delta_t))/(2\delta_t) \quad (26)$$

simply recovers the velocity-Verlet velocities. Higher-order results may be expressed in terms of the velocity-Verlet results at different time steps. For example

$$\mathbf{v}^{(4)}(t) = \frac{1}{6}(8k_1 - k_2) \quad (27)$$

$$\mathbf{v}^{(6)}(t) = \frac{1}{30}(45k_1 - 9k_2 + k_3) \quad (28)$$

$$\mathbf{v}^{(8)}(t) = \frac{1}{420}(672k_1 - 168k_2 + 32k_3 - 3k_4) \quad (29)$$

where

$$k_n = \sum_{m=1}^n \mathbf{v}^{(2)}(t + (2m - n - 1)\delta_t) \quad (30)$$

Rates of change of momenta were calculated by fitting polynomials to successive momenta in an analogous manner. The same procedure may be used in the presence of constraints.

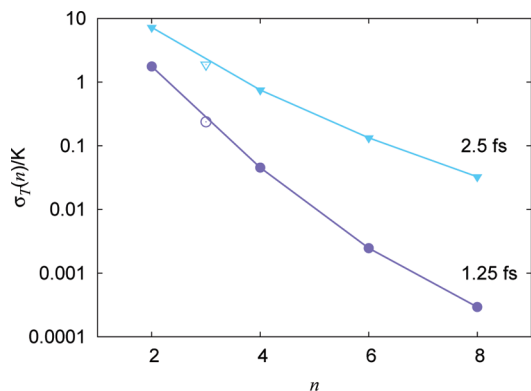


Figure 2. Estimated velocity errors as a function of interpolation order. The weighted rms velocity error defined in eq 31 is given in units of Kelvin. All points are from one of two *NVE* simulations of ubiquitin. The simulation time step for the dark circles was 1.25 fs and for the lighter triangles was 2.5 fs. The lines are just guides to the eye. The solid points are time-symmetric interpolations, and the open points with $n = 3$ are the time-asymmetric (Beeman) interpolations described in the text.

Although the interpolation error in $\mathbf{v}^{(n)}$ is $\mathcal{O}(\delta_t^n)$, there is no guarantee of convergence as n increases. In practice, however, interpolation error appears to rapidly diminish with increasing n , for small n . This is illustrated in Figure 2, where we show the root-mean-square (rms) temperature difference between n^{th} - and tenth-order estimates for protein degrees of freedom in *NVE* simulations of ubiquitin. Specifically, we calculate $\sigma_T(n)$ which we define via

$$\sigma_T^2(n) = \frac{1}{k_B^2 N_f} \sum_{i,\alpha} \langle (p_{i\alpha}(v_{i\alpha}^{(n)} - v_{i\alpha}^{(10)}))^2 \rangle_{\delta_t} \quad (31)$$

where i runs over all protein atoms and α over dimensions x , y , and z . Unless otherwise stated, below we use eighth-order interpolation, which is sufficient to make interpolation error substantially smaller than statistical error in the results we present. Figure 2 suggests that even for the large 2.5 fs time step, eighth-order interpolation corresponds to typical errors in temperature of an individual degree of freedom of less than 0.1 K, and a fourth-order approximation is already a substantial improvement over the second-order result. We have also found interpolation to be useful for trajectories generated by the commonly used r-RESPA method, albeit with slower convergence when tested with outer time steps in the 4–6 fs range (data not shown).

Finally, polynomial fits through positions at times $t + \delta_t$ and earlier are also possible (Figure 2). Although not time symmetric, such interpolations may be useful for thermostats because they have the advantage of giving improved accuracy velocity estimates “on the fly.” In particular, the third-order result recovers the velocities in Beeman’s implementation of Verlet, $\tilde{\mathbf{v}}^{(3)}(t) = (-\mathbf{v}^{(2)}(t - \delta_t) + 2\mathbf{v}^{(1/2)}(t - \delta_t/2) + 2\mathbf{v}^{(2)}(t))/3$, where the half-step velocity is $\mathbf{v}^{(1/2)}(t - \delta_t/2) = (\mathbf{x}(t) - \mathbf{x}(t - \delta_t))/\delta_t$. Beeman’s algorithm is often used instead of velocity Verlet or leapfrog when accurate velocities are important, and Figure 2 shows the improvement over the velocity-Verlet velocities.

4. Results

4.1. Diatomic Fluid. The diatomic fluid is a useful test system, because it contains anharmonicities and motions of different frequencies, yet is simple enough that approximation errors can be estimated based on analytical harmonic oscillator results. We performed velocity-Verlet simulations with different time steps. These ranged up to 5 fs, or about a tenth of the vibrational period of 48 fs. This range was chosen because, in biomolecular simulations, time steps of up to about a tenth of the fastest vibrational period are in common use. (A time step of 2 fs is often chosen, for example, in protein simulations in which bonds to hydrogen are constrained; the fastest motions—angle bending motions involving hydrogens and certain bond stretches—have periods of approximately 20 fs.) For each time step, we calculated the ratio of vibrational and translational temperatures. The translational temperature depends on the center-of-mass velocities and momenta $\{\mathbf{V}, \mathbf{P}\}$, while the vibrational temperature depends on $\{v, p\}$ (the rates of change of the bond lengths $\{r\}$ and the projections of relative momenta $\{\mathbf{p}_r\}$ along the bonds, respectively; see Appendix A):

$$k_B T_{pv}^{\text{trans}} = \frac{N_{\text{mol}}}{N_{\text{mol}} - 1} \overline{\langle \mathbf{P} \cdot \mathbf{V} \rangle}_{\delta_t}, \quad k_B T_{pv}^{\text{vib}} = \overline{\langle p v \rangle}_{\delta_t} \quad (32)$$

The bar denotes an average over all N_{mol} molecules, and the $N_{\text{mol}}/(N_{\text{mol}} - 1)$ prefactor reflects the constraints on the center-of-mass momenta arising from the conservation of the total linear momentum. The atomic velocities were evaluated by polynomial fitting. The center-of-mass and vibrational velocities were calculated in terms of these interpolated velocities. As shown in Figure 3, the ratio $T_{pv}^{\text{vib}}/T_{pv}^{\text{trans}}$ is indeed unity within statistical error, showing that generalized equipartition is satisfied in the simulations.

The same ratio is shown on an enlarged y -scale in Figure 3a and compared to the results obtained using fourth, rather than eighth, order interpolation for the velocities. The observation that generalized equipartition holds, but appears to be violated for larger time steps when the fourth-order estimator is used, is consistent with the theoretical expectation that temperature can be estimated with errors that are smaller (higher order) than $\mathcal{O}(\delta_t^4)$.

We also calculated alternative temperature estimators, starting with the typical p^2 formulas:

$$k_B T_{p^2}^{\text{trans}} = \frac{N_{\text{mol}}/M}{N_{\text{mol}} - 1} \overline{\langle \mathbf{P} \cdot \mathbf{P} \rangle}_{\delta_t}, \quad k_B T_{p^2}^{\text{vib}} = \overline{\langle p^2/\mu \rangle}_{\delta_t} \quad (33)$$

where M and μ are molecular and reduced masses, respectively. As seen in Figure 3, when calculated this way, the vibrational (higher frequency) motion appears to be cooler. In Figure 3, we also show the results of estimating the temperature ratio using the v^2 formulas (with accurate velocity estimates):

$$k_B T_{v^2}^{\text{trans}} = \frac{N_{\text{mol}}/M}{N_{\text{mol}} - 1} \overline{\langle \mathbf{V} \cdot \mathbf{V} \rangle}_{\delta_t}, \quad k_B T_{v^2}^{\text{vib}} = \overline{\langle \mu v^2 \rangle}_{\delta_t} \quad (34)$$

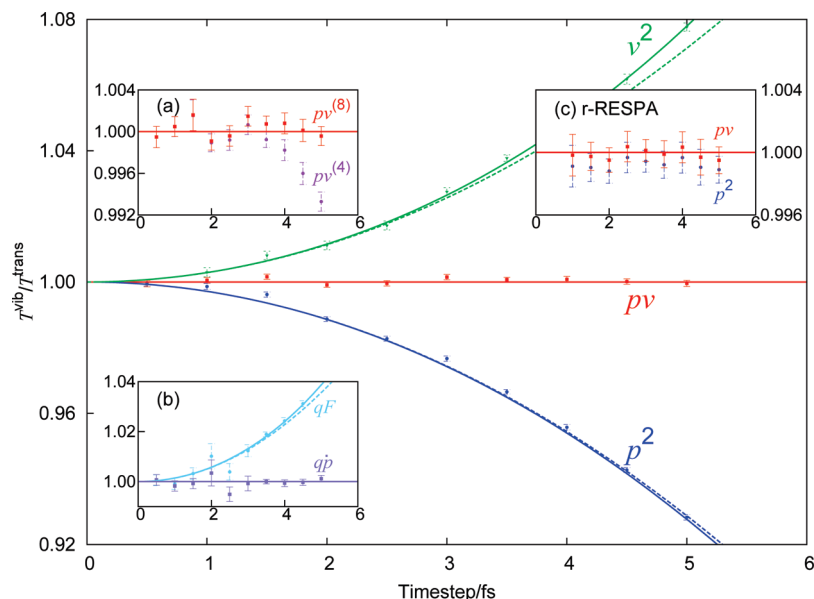


Figure 3. Equipartition for the diatomic fluid over a range of integration time steps. The main figure and insets (a) and (b) show the time-step-dependence of the ratio of different estimates (based on eqs 32–35) of the vibrational and translational temperatures for the same velocity-Verlet simulations. Lines show the harmonic oscillator results (eqs 20–24), with the dashed lines denoting asymptotes that only include terms up to $(\omega_0\delta_t)^2$. The main figure shows that, with the pv estimator, the ratio $T^{\text{vib}}/T^{\text{trans}}$ is one within error, indicating that generalized equipartition is achieved in the simulations. The p^2 estimator (blue) gives a ratio less than one, making the vibrational motions appear cooler, whereas the v^2 estimator gives a ratio greater than one. Inset (a) shows only the results of the pv estimator in a region close to $T^{\text{vib}}/T^{\text{trans}} = 1$. The red points are identical to those in the main figure, and the purple circles show the result of using velocities obtained from a lower-order polynomial interpolation (fourth rather than eighth). Inset (b) shows that the qp estimator gives a temperature ratio close to one, confirming generalized equipartition is satisfied, but that the qF estimator (see main text) makes the vibrations appear hotter. Inset (c) shows the result of the pv and p^2 estimators for r-RESPA simulations (with a fixed inner time step of 0.5 fs) as a function of the outer time step.

In this case, the higher frequency motion appears hotter. The magnitude of the deviations from conventional equipartition that are revealed using eqs 33 and 34 are also seen to be well predicted using the harmonic oscillator results of Section 2.4.

As shown in Appendix A, T_{qp}^{vib} , defined by

$$-k_{\text{B}}T_{qp}^{\text{vib}} = \overline{\langle r\dot{p} \rangle_{\delta_t}} \quad (35)$$

is also essentially equal to the vibrational temperature. We calculated T_{qp}^{vib} in terms of the bond vectors $\{\mathbf{r}\}$ and the molecular angular momenta and velocities $\{\mathbf{L}, \boldsymbol{\omega}\}$ using the identity $r\dot{p} \equiv \mathbf{r} \cdot \dot{\mathbf{p}}_{\mathbf{r}} + \mathbf{L} \cdot \boldsymbol{\omega}$, with $\mathbf{p}_{\mathbf{r}}$ and the atomic velocities contained in $\boldsymbol{\omega}$ obtained from interpolation. Additionally, we calculated the analogous, but approximate, temperature estimator T_{qF}^{vib} (Appendix B). Figure 3b shows that T_{qp}^{vib} is the same as the translational temperature within statistical error, consistent with generalized equipartition being satisfied. In contrast, the T_{qF}^{vib} estimator leads to hotter temperatures for the vibrational (i.e., higher frequency) motion. This deviation can again be understood in terms of the harmonic oscillator results. T_{qF} is relevant to pressure computations, as described in Appendix B. We find that the average pressure calculated using a p^2 expression for the ideal part and a virial part calculated in the usual manner¹³ depends on whether a molecular or atomic expression is used. For the smallest (0.5 fs) time step, the molecular and atomic results are the same within statistical error of 0.5 bar, but for the largest (5 fs) time step they differ by 17.8 ± 0.4 bar, which is close to the 17.0 bar predicted by eq 44.

Although eqs 33 and 34 are less accurate than eq 32 for estimating temperature, the deviations from conventional equipartition that they reveal do reflect real differences between the shadow Hamiltonian and H_0 . They thus provide a warning that truncation errors may affect other quantities. We find vibrational frequencies and the magnitude of bond length fluctuations to change, for example, by approximately 2% over the range of time steps studied, as would be expected on the basis of the harmonic oscillator results. One straightforward way to reduce truncation errors is to use a multiple-time-step scheme, where the stiff-bonded forces are evaluated more frequently than the softer intermolecular interactions. Since the intermolecular interactions usually dominate the computational expense, this approach often only has modest cost. Figure 3c shows the vibrational to translational temperature ratios for the pv and p^2 estimators from r-RESPA simulations of the diatomic fluid as a function of the outer time step. The intermolecular interactions were calculated on the outer time step, while the bonded interactions were evaluated every 0.5 fs. The agreement between the estimators is excellent even for large outer time steps; the v^2 estimator (not shown) is also in agreement with the other estimators.

4.2. Ubiquitin. As described in Appendix A, it is straightforward to calculate the temperature of different quasiharmonic motions of a protein. We have done this for different simulations of ubiquitin solvated in water. Initially, we performed *NVE* simulations with differing values of δ_t . We calculated the temperature $T_{pv}^{(i)}$ using eq 40 for all quasihar-

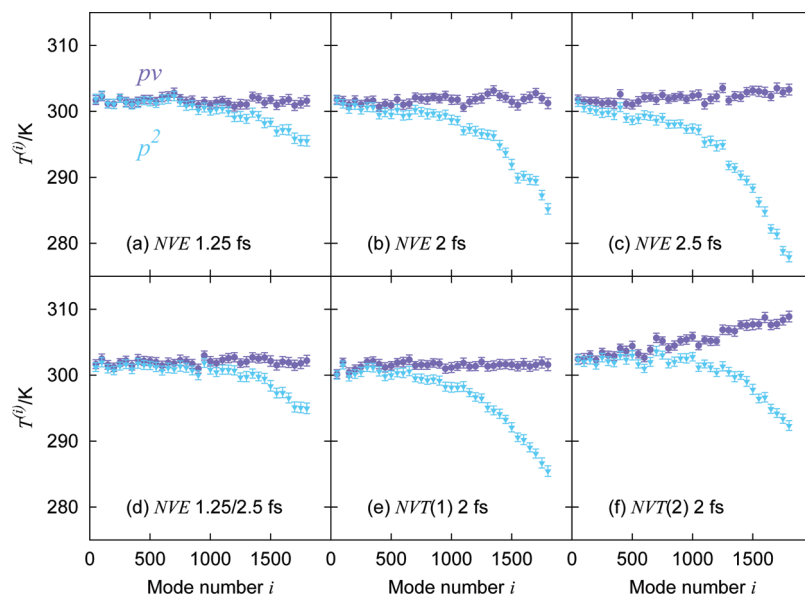


Figure 4. Temperature as a function of mode number i for ubiquitin. The modes are ordered by mean-square fluctuation, with the largest-amplitude (lowest-frequency) motions to the left. The temperatures are averages over 50 consecutive modes. The dark circles show the results of the pv formula (eq 40), and the light triangles are the results of the corresponding p^2 formula. Panels (a–c) show the results of NVE simulations with time steps of 1.25, 2, and 2.5 fs. Panel (d) shows results of the NVE simulation using a r-RESPA multiple time step method. Panel (e) shows the results when the entire system is coupled to a single thermostat, whereas panel (f) shows the results of simultaneously applying independent thermostats to water and protein.

monic modes i . As shown in Figure 4a–c, which corresponds to time steps of 1.25, 2, and 2.5 fs, $T_{pv}^{(i)}$ is essentially independent of i . This strongly suggests that generalized equipartition is indeed achieved in these simulations. We also estimated temperatures using $k_B T_{p^2}^{(i)} = \langle p_i^2 \rangle_{\delta}$. This estimate varies substantially by mode and, as expected, is significantly lower for higher-frequency motions. The deviation from conventional equipartition increases with the size of the time step (Figure 4). We also find $T_{v^2}^{(i)} = \langle v_i^2 \rangle_{\delta} / k_B$ to have a strong but opposite dependence on mode (not shown). Note that all three estimates agree well for the low-frequency motions. Since the v^2 formulas are found to overestimate temperature by approximately as much as the p^2 formula underestimates it, we just discuss the pv and p^2 results below.

It is natural to try to reduce the truncation errors signaled by the breakdown of conventional equipartition. As noted above, one approach is to use a multiple-time-step method. In Figure 4d, we show the results of using the r-RESPA method with bonded forces calculated every 1.25 fs and with nonbonded interactions every 2.5 fs. The magnitude of the discrepancy between $T_{pv}^{(i)}$ and $T_{pp}^{(i)}$ is reduced to an amount similar to the 1.25 fs time step velocity-Verlet simulation. This is because the highest frequency motions are bond vibrations and angle-bending motions that involve hydrogen atoms. The r-RESPA solution is inexpensive, if calculating the bonded interactions takes a relatively small part of the overall computation time, which is typically the case.

If a thermostat is applied to the system and the instantaneous temperature—which determines thermostat, and hence particle, dynamics—is estimated using the p^2 formula, then integration errors may be amplified. A common, if fairly innocuous, case is illustrated in Figure 4e, where a single Berendsen thermostat is applied to the entire system. The

Table 1. Water-Protein Temperature Differences from p^2 and v^2 Estimators for NVE Simulations^a

time step/fs	ΔT_{p^2}	ΔT_{pv}
1.25	2.11	−0.11
2.0	5.62	−0.06
2.5	8.91	−0.08

^a Temperature differences are expressed as $\Delta T = T^{\text{water}} - T^{\text{protein}}$ in Kelvin. Statistical errors are approximately 0.1 to 0.2 K.

temperature as a function of mode is very similar to the NVE simulation with the same (2 fs) time step. There is a small discrepancy between the system temperature (301.76 ± 0.02 K) and the target temperature of 300 K. The discrepancy reflects the fact that the p^2 formula slightly underestimates the water temperature ($T_{p^2}^{\text{water}} < T_{pv}^{\text{water}}$).

Water models used in biomolecular simulation are often rigid, so the highest frequency motions are in the protein. Thus, although with the p^2 formula the water may only appear a degree or two cooler than the true value, the protein may appear substantially cooler, leading to an apparent temperature difference, $\Delta T_{p^2} \equiv T_{p^2}^{\text{water}} - T_{p^2}^{\text{protein}}$. In the NVE simulations with a 2 fs time step, for example, $\Delta T_{p^2} = 5.6 \pm 0.1$ K, and this rises to 8.9 ± 0.2 K for a 2.5 fs time step, whereas the accurately calculated temperature difference $\Delta T_{pv} \equiv T_{pv}^{\text{water}} - T_{pv}^{\text{protein}}$ is 0 within error in both cases; see Table 1). If the symptom of truncation error revealed by the p^2 estimator is combatted by applying two thermostats simultaneously—one to protein and one to solvent—then larger errors result than in the case of a single system-wide thermostat. This is not due to an intrinsic problem with the use of multiple thermostats, which can be used safely if an appropriate time step is chosen. Rather, the relatively large time steps commonly chosen in MD simulations for ef-

Table 2. Estimated Component Temperatures for NVT Simulations Performed with Different Thermostats^a

thermostat	$T_{p^2}^{\text{system}}$	$T_{p^2}^{\text{water}}$	$T_{p^2}^{\text{protein}}$	T_{pv}^{system}	T_{pv}^{water}	T_{pv}^{protein}
one (system)	299.97	300.47	294.80	301.76	301.76	301.78
one (water)	299.49	299.97	294.55	301.28	301.26	301.52
one (protein)	305.02	305.50	300.03	306.84	306.81	307.14
two (protein, water)	299.98	300.05	299.18	301.78	301.35	306.30

^a Temperatures are in Kelvin. The target temperature was 300.0 K in all simulations. The simulations differ only in the number of thermostats (one or two) and the atoms to which the thermostats are coupled. Statistical errors are approximately 0.1 to 0.2 K for protein temperatures and 0.02 to 0.04 K for water/system temperatures.

iciency lead to errors that can be amplified by certain choices of thermostat. For our simulation with a 2 fs time step and two thermostats, we find that $T_{pv}^{\text{protein}} = 306.3 \pm 0.1$ K, whereas $T_{pv}^{\text{water}} = 301.35 \pm 0.02$ K; in addition $T_{pv}^{(i)}$ is no longer approximately constant (Figure 4f). This signals a breakdown of generalized equipartition. Energy flows from one thermostat to the protein, then transfers to the water, and is finally removed by the other thermostat. Applying a single thermostat to the system, but just coupling it to a subset of the particles, should not lead to a breakdown of generalized equipartition, although if the component contains high-frequency motions, then this can lead to a substantial error in the system temperature. For example, we find that coupling a single thermostat to the protein leads to a simulation temperature about 7 K above the target temperature; see Table 2. We stress that the underlying cause of these problems is truncation error, not a problem with the thermostat itself.

5. Discussion and Conclusions

In this paper, using the established concepts of generalized equipartition and the shadow Hamiltonian, we have introduced a clear definition of simulation temperature that explicitly takes into account the finite simulation time step. We have shown that this temperature can be evaluated accurately and straightforwardly in practice. We tested generalized equipartition in numerical examples relevant to biomolecular simulation in which truncation errors lead to deviations from conventional equipartition and thus to different temperature estimates for different motions when conventional estimators are used. We confirmed that generalized equipartition is in fact satisfied in these examples, with different motions sharing a single well-defined simulation temperature.

The observation that generalized equipartition can be satisfied even for rather large time steps naturally does not imply that the simulations are free from artifacts due to truncation error, but it does help highlight the actual nature of the errors. As signaled by the breakdown of conventional equipartition, the shadow Hamiltonian differs from the Hamiltonian that we wish to simulate by an amount $\mathcal{O}(\delta_i^2)$, and thus their dynamics and thermodynamics will differ too.

One practical benefit of obtaining accurate temperature estimates, even when the Hamiltonian itself is subject to $\mathcal{O}(\delta_i^2)$ errors, is that testing generalized equipartition can be

a valuable simulation diagnostic. A violation of equipartition demonstrated using the methods of this paper points to underlying problems with the integration scheme, as in the two-thermostat example described above. A second practical benefit is that accurately estimating the temperature can remove what may be the largest source of error in the description of low-frequency motions, as we now briefly explain. Low-frequency motions are often of greatest interest, and by their nature, most error in their description comes via their coupling to higher-frequency motions, which present more of a challenge to the integrator. If high- and low-frequency motions are weakly coupled, as expected for bond vibrations and larger-scale protein conformational change, for example, the low-frequency dynamics should be accurately described by the integrator. Error in the estimated temperature can then become the dominant error in the overall description of the low-frequency motion, because the temperature—when computed as a sum over all atomic motions using the p^2 formula—is polluted by errors due to the fast motions.

Our results also make clear that for Verlet integration, estimating the temperature using the v^2 formula leads to $\mathcal{O}(\delta_i^2)$ errors even if the velocities could be computed exactly. A corollary is that Beeman's version of Verlet, which gives velocities with only $\mathcal{O}(\delta_i^3)$ errors, will still yield temperatures with $\mathcal{O}(\delta_i^2)$ errors if those temperatures are estimated using a v^2 formula, as is conventional when using this integrator. Most simulations use some form of temperature control, and an inaccurate estimated temperature can affect the dynamics through the thermostat (or barostat). Fortunately, for the common case of a small globular protein solvated by constrained water molecules and coupled to a single system-wide thermostat, the resultant errors will be small because the fastest motions are in the protein, which comprises only a small part of the system. Care might be needed if the system contains a larger fraction of high-frequency motions, as would be the case in a simulation of a protein crystal or a lipid bilayer or in a simulation using an unconstrained water model. Clearly, systems that are particularly sensitive to temperature and pressure are more likely to exhibit substantial artifacts. Systems near a phase transition, for example, need more care; under ambient conditions, such systems include certain lipid bilayers and marginally stable small peptides and proteins.

For the simulation thermostat to accurately control temperature, it would be desirable to calculate the instantaneous temperature using the pv formula. We have shown that improved estimates of temperature at a given time are possible using information that can in principle be made available by the integrator (see the non-time-symmetric interpolation in Figure 2). Although constructing a thermostat along these lines is possible, this may not be the most promising approach. In addition to breaking time reversibility, such a thermostat would require the MD code to retain information about particle positions from earlier time steps, thereby adding complexity and likely reducing performance of a parallel code. One simple way to side-step these issues may be to continue to use the p^2 formula but to couple the thermostat to lower-frequency motions. This approach does

not remove $\mathcal{O}(\delta_t^2)$ errors from the p^2 temperature estimate, but it can substantially reduce the prefactor. In our *NVE* ubiquitin simulations, for example, we find the error in $T_{p^2}^{\text{water}}$ to be reduced by a factor of over five when it is calculated from translational motion only, rather than translational and rotational. This suggests that coupling a single thermostat to the translational motion of water molecules may be a useful approach.

Our results warn against combining multiple thermostats with large time steps. With a properly chosen time step, multiple thermostats can be a valuable tool to ensure equilibration even when equipartitioning is slow.⁴² Application of multiple thermostats was also once useful to control very large hot solvent–cold solute artifacts, such as can occur when cutoff electrostatics are used. The improvements in methodology and computer codes over the last 15–20 years, however, have led to a situation where the dominant deviations from conventional equipartition are truncation errors and the generalized equipartition is satisfied. In such a situation, using multiple thermostats to rectify the deviation from conventional equipartition will have the counterproductive effect of causing true temperature imbalances in the system. Although in this paper we reported results obtained with the Berendsen thermostat, we have also found very similar results in tests with a Nosé–Hoover thermostat. Caution may also be required when combining stochastic thermostats with large time steps (particularly if the thermostat relaxation time is short), because such thermostats are typically coupled to many individual degrees of freedom (as in Langevin dynamics, for example).

The simplest way to reduce truncation error is obvious: reduce the time step. In practice, there is often reluctance to do this, in part because of the large computational expense of simulations and in part because of the fact that, while artifacts undeniably exist, their direct impact is largest on fast motions and their effect on properties likely to be of interest in long-time-scale simulations is much less clear. Indeed, partly motivated by the observation that even the large commonly used time steps (of about a tenth of the period of the fastest motions) are approximately a factor of three below the stability limit of velocity Verlet, some authors have suggested increasing the time step further.^{43,44} By showing that generalized equipartition can hold even for time steps somewhat beyond the commonly used range, our results lend some support to this idea. On the other hand, regardless of whether generalized equipartition is satisfied, truncation errors will affect simulation results, and it is difficult to assess the impact on properties of interest in complicated biological systems. Thus, a more promising approach to balancing accuracy and efficiency may be to change the integrator. Results for both our test systems highlighted the effectiveness of the r-RESPA integrator for reducing errors at little cost; such an approach is likely to be useful for biomolecular simulation, since the fast motions that are the major source of truncation error are usually inexpensive to calculate and can thus be calculated with a reduced time step at little cost. In some cases, in particular on specialized hardware that greatly accelerates nonbonded interactions,⁴⁵ a substantial fraction of time may be spent on bonded interactions. In

future work, we will describe new integrators that increase accuracy efficiently in such cases.

Appendix A

Alternative Coordinate Systems. It can be helpful to calculate temperatures for different modes of motion, such as the collective motions of a large subset of atoms. In principle, a straightforward recipe to do this is to identify a canonical transformation between the Cartesian atomic coordinates and conjugate momenta and a set of coordinates of interest. If this can be done, generalized equipartition should then hold for the new variables, which can be evaluated in terms of the atomic coordinates and momenta provided by the integrator. Time derivatives of the generalized coordinates may be obtained using interpolation. We illustrate with two examples relevant to the systems studied in this paper.

Translation, Vibration, and Rotation for a Diatomic Molecule. Consider a diatomic molecule, which may be part of a larger system, that consists of atoms A and B with mass m_A and m_B , respectively. The transformation from the atomic positions and momenta ($\mathbf{q}_A, \mathbf{q}_B, \mathbf{p}_A, \mathbf{p}_B$) to center-of-mass and relative positions and momenta ($\mathbf{R} = (m_A\mathbf{q}_A + m_B\mathbf{q}_B)/M$, $\mathbf{r} \equiv (x, y, z) = \mathbf{q}_B - \mathbf{q}_A$, $\mathbf{P} = \mathbf{p}_A + \mathbf{p}_B$, $\mathbf{p} \equiv (p_{rx}, p_{ry}, p_{rz}) = \mu(\mathbf{p}_B/m_B - \mathbf{p}_A/m_A)$), where $M = m_A + m_B$ and $1/\mu = 1/m_A + 1/m_B$, is canonical. So is the further transformation of the relative motion into vibrational and rotational motion ($r = |\mathbf{r}|$, $\theta = \arccos(z/r)$, $\phi = \arctan(y/x)$, $p = \mathbf{p}_r \cdot \mathbf{r}/r$, $l_\theta = -(x^2 + y^2)^{1/2}p_{rz} + (xp_{rx} + yp_{ry})z/(x^2 + y^2)^{1/2}$, $l_\phi = xp_{ry} - yp_{rx}$). Generalized equipartition relations may thus be written for translational and internal motion and for rotational and vibrational contributions to the internal motion:

$$\frac{1}{3}\langle \mathbf{P} \cdot \mathbf{V} \rangle_{\delta_t} = \frac{1}{3}\langle \mathbf{p}_r \cdot \mathbf{v}_r \rangle_{\delta_t} = k_B T \quad (36)$$

$$\frac{1}{2}\langle \mathbf{L} \cdot \boldsymbol{\omega} \rangle_{\delta_t} = \langle pv \rangle_{\delta_t} = k_B T \quad (37)$$

Here $\mathbf{V} = \dot{\mathbf{R}}$, $\mathbf{v}_r = \dot{\mathbf{r}}$, $v = \dot{r}$, and we have identified $\mathbf{L} \cdot \boldsymbol{\omega} \equiv l_\theta \dot{\theta} + l_\phi \dot{\phi}$, where by definition $\mathbf{L} = \mathbf{r} \times \mathbf{p}_r$ and $\boldsymbol{\omega} = \mathbf{r} \times \mathbf{v}_r/r^2$ as usual. Likewise, using eq 3, we have the additional expressions for the total internal and the vibrational motion:

$$\frac{1}{3}\langle \mathbf{r} \cdot \dot{\mathbf{p}}_r \rangle_{\delta_t} = \langle rp \dot{p} \rangle_{\delta_t} = -k_B T \quad (38)$$

which are valid even with periodic boundary conditions (in the unlikely situation that the bond length can exceed half the simulation box length, care must be taken not to incorrectly wrap the relative position coordinate). Replacement of $\dot{\mathbf{p}}_r$ in the above formula with the analogously defined relative force

$$\mathbf{F}_r = \mu(\mathbf{F}_B/m_B - \mathbf{F}_A/m_A) \quad (39)$$

where $\mathbf{F}_A = -\partial U/\partial \mathbf{q}_A$ denotes an atomic force, would be an approximation for finite integration time steps.

The generalized equipartition formulas obtained in this section are very familiar in the case $\delta_t \rightarrow 0$; the formal reasoning here makes clear that they should also apply to

the shadow Hamiltonian and thus hold in simulation, provided that the time derivatives that they contain can be accurately estimated.

Quasiharmonic Motions of a Protein. Quasiharmonic analysis and principal component analysis are popular closely related methods for analyzing protein motions⁴⁶ and have occasionally been used in the context of equipartition.⁴⁷ The basic approach can often be decomposed into three transformations: First, and optionally, the protein atoms or some subset are mass-weighted ($\mathbf{q}_0, \mathbf{p}_0 \rightarrow \mathbf{q}_1 = \mathbf{m}^{1/2}\mathbf{q}, \mathbf{p}_1 = \mathbf{m}^{-1/2}\mathbf{p}$); then these coordinates are transformed by means of a \mathbf{q}_1 -dependent overall translation and rotation to minimize the root-mean-square deviation (rmsd) of \mathbf{q}_1 to a reference structure, ($\mathbf{q}_1, \mathbf{p}_1 \rightarrow \mathbf{q}_2 = \mathbf{A}_{\text{rot}}\hat{\mathbf{A}}_{\text{trans}}\mathbf{q}_1, \mathbf{p}_2 = \mathbf{A}_{\text{rot}}\mathbf{p}_1$); finally, an orthogonal transformation ($\mathbf{q}_2, \mathbf{p}_2 \rightarrow \mathbf{q}' = \mathbf{R}^T\mathbf{q}_2, \mathbf{p}' = \mathbf{R}^T\mathbf{p}_2$) makes the covariance matrix $\langle(\mathbf{q}' - \langle\mathbf{q}'\rangle)(\mathbf{q}' - \langle\mathbf{q}'\rangle)^T\rangle_{\delta_i} = \mathbf{R}^T\langle(\mathbf{q}_2 - \langle\mathbf{q}_2\rangle)(\mathbf{q}_2 - \langle\mathbf{q}_2\rangle)^T\rangle_{\delta_i}\mathbf{R}$ diagonal. Motion along a subset of the \mathbf{q}' coordinates with the largest eigenvalues often correspond to interesting fluctuations of the protein around its native state. It is natural to define a temperature for mode i via

$$\langle p'_i v'_i \rangle_{\delta_i} = k_B T_{pv}^{(i)} \quad (40)$$

(or $\langle q'_i p'_i \rangle_{\delta_i} = k_B T_{qp}^{(i)}$), where $v'_i = \dot{q}'_i$. Unfortunately, although the first and last transformations above are canonical, the momenta generated in the rmsd-fitting step are only approximations to the true conjugate momenta. We thus expect $T_{pv}^{(i)}$ to differ slightly from T . In practice, we expect this discrepancy to be small for an ordered protein with a large number of degrees of freedom, and we neglect it.

Appendix B

Pressure. We show here, by means of a simple example, that the equivalence of atomic and molecular definitions of the simulation pressure for $\delta_i > 0$ may be viewed as a consequence of generalized equipartition and that this equivalence is broken, if the virial is computed in the normal way and the temperature is evaluated using a p^2 (or indeed pv or v^2) estimate. Although it is straightforward to obtain exact expressions for the simulation pressure, we have not found a practical method to estimate it from simulation data in a way that is as simple as evaluating the simulation temperature. Our results suggest that using a pv estimate for temperature will reduce pressure errors relative to a p^2 estimate but not eliminate them. This is consistent with the work of Pastor et al., who demonstrated that a different estimator of temperature (derived from Verlet velocities from the previous half-step) leads to exact estimates of pressure for a harmonic oscillator (unlike the use of p^2 , pv , or v^2 estimates) due to a favorable cancellation of errors.⁴⁸

In the canonical ensemble, starting from the thermodynamic definition of pressure as a volume derivative of the free energy, $P = -\partial F/\partial V$, it is straightforward to express the simulation pressure as a sum of kinetic and virial contributions. Assuming periodic boundary conditions and no constraints and viewing the Hamiltonian as a function of atomic positions and momenta, one obtains the *atomic* expression:

$$P = \frac{N}{V}k_B T - \left\langle \left(\frac{\partial H_{\delta_i}(\mathbf{p}, \mathbf{q}, V)}{\partial V} \right)_{\mathbf{p}, \mathbf{q}} \right\rangle_{\delta_i} \quad (41)$$

In terms of molecular center-of-mass positions and momenta $\{\mathbf{R}, \mathbf{P}\}$, and relative coordinates $\{\mathbf{r}, \mathbf{p}_r\}$, the following *molecular* expression is more natural:

$$P = \frac{N_{\text{mol}}}{V}k_B T - \left\langle \left(\frac{\partial H_{\delta_i}(\{\mathbf{P}, \mathbf{R}, \mathbf{p}_r, \mathbf{r}\}, V)}{\partial V} \right)_{\{\mathbf{P}, \mathbf{R}, \mathbf{p}_r, \mathbf{r}\}} \right\rangle_{\delta_i} \quad (42)$$

N_{mol} denotes the number of molecules, and N denotes the number of atoms. The two expressions are equivalent, but this equivalence can be broken by the (approximate) method used to compute the temperature and virial, as we explain by means of a simple example.

Consider an ideal diatomic gas with some intramolecular bonded interaction but negligible intermolecular interactions. Since the molecular virial vanishes, the molecular pressure formula immediately yields the correct result, $PV = N_{\text{mol}}k_B T$. (Assuming that the T appearing in the molecular pressure is obtained from the kinetic energy of molecular center-of-mass motion, T will be estimated correctly for this idealized system regardless of whether a pv or p^2 formula is used, provided the integrator—like velocity Verlet—preserves translational invariance.) We find that the atomic virial of our ideal system reduces to $(N_{\text{mol}}/3)\langle \mathbf{r} \cdot \dot{\mathbf{p}}_r \rangle_{\delta_i}$, where the overbar simply denotes an average over all molecules. If generalized equipartition (eq 38) holds, then the atomic virial further reduces to $-N_{\text{mol}}k_B T$ and thus precisely cancels half of the kinetic term, yielding the correct pressure. The natural approach to computing the pressure when using velocity-Verlet integration, however, is to estimate T from the atomic momenta using a p^2 formula and from the atomic virial using $(N_{\text{mol}}/3)\langle \mathbf{r} \cdot (-\partial U/\partial \mathbf{r}) \rangle_{\delta_i}$. (This is essentially the approach implemented in Desmond,³² for example.) Then we find that the atomic pressure differs from the molecular pressure according to

$$(P_{\text{mol}}^{\text{ideal}} - P_{\text{atom}}^{\text{ideal}})V = \frac{1}{3}N_{\text{mol}}k_B(T_{qF}^{\text{vib}} - T_{p^2}^{\text{vib}}) \quad (43)$$

where the two different approximations to the vibrational temperatures are $T_{p^2}^{\text{vib}} = \langle p^2/\mu \rangle_{\delta_i}$, and $T_{qF}^{\text{vib}} = \langle rF_{\text{eff}} \rangle_{\delta_i}$. The effective force includes a centrifugal term and is defined via $rF_{\text{eff}} = \mathbf{r} \cdot \mathbf{F}_r + \mathbf{L} \cdot \mathbf{L}/(\mu r^2)$, with \mathbf{F}_r defined in eq 39.

Equation 43 depends on the details of the intramolecular interaction, but if we assume that the effective intramolecular potential, i.e., with a $L^2/(2\mu r^2)$ centrifugal term included, is approximately harmonic, then we may use the results of Section 2.4 to yield

$$\begin{aligned} (P_{\text{mol}}^{\text{ideal}} - P_{\text{atom}}^{\text{ideal}})V &= \frac{1}{3}N_{\text{mol}}k_B T \left(\frac{k}{k_{\delta_i}} - \frac{m_{\delta_i}}{m} \right) \\ &\approx \frac{1}{3}N_{\text{mol}}k_B T \left(\frac{1}{12}(\omega\delta_i)^2 + \frac{1}{6}(\omega\delta_i)^2 \right) \\ &= N_{\text{mol}}k_B T \frac{(\omega\delta_i)^2}{12} \end{aligned} \quad (44)$$

Using the atomic formula, in the way described above, will thus underestimate the simulation pressure. For a time step

of $\omega\delta_t = 0.6$, the overall error in the atomic pressure is 3% of the ideal gas value. For condensed-phase systems under ambient conditions, similar errors could easily dominate the pressure, which is itself the difference of two large almost canceling terms. For compressible systems, such as membranes, this is cause for caution. Using a molecular (or group-based) pressure is one obvious way to reduce errors (in our simple example, this approach eliminates errors). Alternatively, or in addition, the temperature estimate could be improved. Since two-thirds of the error in the atomic pressure estimate originates in the temperature estimate and one-third from the virial estimate, use of a perfect estimate of temperature will improve the pressure estimate but will not eliminate errors. As noted above, however, Pastor et al. have shown how the temperature estimate may be changed to cancel errors arising from the virial part, thus providing an even more accurate pressure estimator; this method is available in CHARMM.²⁷

Acknowledgment. We thank Rebecca Kastleman for editorial assistance.

References

- Huang, K. *Statistical Mechanics*; John Wiley & Sons: New York, 1987.
- Straub, J. E.; Thirumalai, D. *Proteins: Struct., Funct., Genet.* **1993**, *15*, 360–373.
- Tobias, D. J.; Martyna, G. J.; Klein, M. L. *J. Phys. Chem.* **1993**, *97*, 12959–12966.
- Lingenheil, M.; Denschlag, R.; Reichold, R.; Tavan, P. *J. Chem. Theory Comput.* **2008**, *4*, 1293–1306.
- Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.
- Harvey, S. C.; Tan, R. K.-Z.; Cheatham, T. E., III. *J. Comput. Chem.* **1998**, *19*, 726–740.
- Mor, A.; Ziv, G.; Levy, Y. *J. Comput. Chem.* **2008**, *29*, 1992–1998.
- Levitt, M.; Sharon, R. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 7557–7561.
- Guenot, J.; Kollman, P. A. *Protein Sci.* **1992**, *1*, 1185–1205.
- Arnold, G. E.; Ornstein, R. L. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 19–33.
- Sagui, C.; Darden, T. A. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155–179.
- Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637–649.
- Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Oxford University Press: New York, 1989.
- Ruth, R. D. *IEEE Trans. Nucl. Sci.* **1983**, *30*, 2669–2671.
- Hairer, E.; Lubich, C.; Wanner, G. Geometric Numerical Integration. In *Structure-Preserving Algorithms for Ordinary Differential Equations; volume 31 of Springer Series in Computational Mathematics*; Springer: Berlin, Germany, 2006.
- Tolman, R. C. *Phys. Rev.* **1918**, *11*, 261–275.
- Gans, J.; Shalloway, D. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2000**, *61*, 4587–4592.
- Cuendet, M. A.; van Gunsteren, W. F. *J. Chem. Phys.* **2007**, *127*, 184102.
- MacGowan, D.; Heyes, D. M. *Mol. Simul.* **1988**, *1*, 277–297.
- Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990–2001.
- Uline, M. J.; Siderus, D. W.; Corti, D. S. *J. Chem. Phys.* **2008**, *128*, 124301.
- Shirts, R. B.; Burt, S. R.; Johnson, A. M. *J. Chem. Phys.* **2006**, *125*, 164102.
- Leimkuhler, B. J.; Skeel, R. D. *J. Comput. Phys.* **1994**, *112*, 117–125.
- Bond, S. D.; Leimkuhler, B. J. *Acta Numerica* **2007**, *16*, 1–65.
- Engle, R. D.; Skeel, R. D.; Drees, M. J. *J. Comput. Phys.* **2005**, *206*, 432–452.
- Zhou, J.; Reich, S.; Brooks, B. R. *J. Chem. Phys.* **2000**, *112*, 7919–7929.
- Brooks, B. R. et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
- Izaguirre, J. A.; Hampton, S. S. *J. Comput. Phys.* **2004**, *200*, 581–604.
- Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695–1697.
- Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossváry, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Novel Algorithms for Scalable Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06)*; ACM/IEEE Conference on Supercomputing (SC06), Tampa, FL, November 11–17, 2006; ACM Press: New York, 2006.
- Flyvbjerg, H.; Petersen, H. G. *J. Chem. Phys.* **1989**, *91*, 461–466.
- Kaminski, G.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
- Maragakis, P.; Lindorff-Larsen, K.; Eastwood, M. P.; Dror, R. O.; Klepeis, J. L.; Arkin, I. T.; Jensen, M. Ø.; Xu, H.; Trbovic, N.; Friesner, R. A.; Palmer, A. G., III.; Shaw, D. E. *J. Phys. Chem. B* **2008**, *112*, 6155–6158.
- Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.
- Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Mod.* **2001**, *7*, 306–317.
- Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981.
- Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- Kräutler, V.; Van Gunsteren, W. F.; Hünenberger, P. H. *J. Comput. Chem.* **2001**, *22*, 501–508.

- (41) Lippert, R. A.; Bowers, K. J.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossváry, I.; Shaw, D. E. *J. Chem. Phys.* **2007**, *126*, 046101.
- (42) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J.; Klein, M. L. *J. Chem. Phys.* **1993**, *99*, 2796–2808.
- (43) Mazur, A. K. *J. Comput. Phys.* **1997**, *136*, 354–365.
- (44) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786–798.
- (45) Shaw, D. E. et al. Anton, a special-purpose machine for molecular dynamics simulation. In *Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA '07)*; 34th Annual International Symposium on Computer Architecture (ISCA '07), San Diego, CA, June 9–13, 2007ACM Press: New York, 2007.
- (46) Case, D. A. *Curr. Opin. Struct. Biol.* **1994**, *4*, 285–290.
- (47) Oda, K.; Miyagawa, H.; Kitamura, K. *Mol. Simul.* **1996**, *16*, 167–177.
- (48) Pastor, R. W.; Brooks, B. R.; Szabo, A. *Mol. Phys.* **1988**, *65*, 1409–1419.
CT9002916

Polarizable Molecular Dynamics Simulation of Zn(II) in Water Using the AMOEBA Force Field

Johnny C. Wu,[†] Jean-Philip Piquemal,^{*,‡,§} Robin Chaudret,^{‡,§} Peter Reinhardt,^{‡,§} and Pengyu Ren^{*,†}

Department of Biomedical Engineering, University of Texas at Austin, Austin, Texas 78712-1062, UPMC Univ. Paris 06, UMR 7616, Laboratoire de Chimie Théorique, case courrier 137, 4 place Jussieu, F-75005, Paris, France, and CNRS, UMR 7616, Laboratoire de Chimie Théorique, case courrier 137, 4 place Jussieu, F-75005, Paris, France

Received February 15, 2010

Abstract: The hydration free energy, structure, and dynamics of the zinc divalent cation are studied using a polarizable force field in molecular dynamics simulations. Parameters for the Zn^{2+} are derived from gas-phase *ab initio* calculation of the Zn^{2+} –water dimer. The Thole-based dipole polarization is adjusted on the basis of the constrained space orbital variations (CSOV) calculation, while the symmetry adapted perturbation theory (SAPT) approach is also discussed. The vdW parameters of Zn^{2+} have been obtained by comparing the AMOEBA Zn^{2+} –water dimerization energy with results from several theory levels and basis sets over a range of distances. Molecular dynamics simulations of Zn^{2+} solvation in bulk water are subsequently performed with the polarizable force field. The calculated first-shell water coordination number, water residence time, and free energy of hydration are consistent with experimental and previous theoretical values. The study is supplemented with extensive reduced variational space (RVS) and electron localization function (ELF) computations in order to unravel the nature of the bonding in $\text{Zn}^{2+}(\text{H}_2\text{O})_n$ ($n = 1, 6$) complexes and to analyze the charge transfer contribution to the complexes. Results show that the importance of charge transfer decreases as the size of the Zn–water cluster grows due to anticooperativity and to changes in the nature of the metal–ligand bonds. Induction could be dominated by polarization when the system approaches the condensed phase and the covalent effects are eliminated from the Zn(II)–water interaction. To construct an “effective” classical polarizable potential for Zn^{2+} in bulk water, one should therefore avoid overfitting to the *ab initio* charge transfer energy of the Zn^{2+} –water dimer. Indeed, in order to avoid overestimation of the condensed-phase many-body effects, which is crucial to the transferability of polarizable molecular dynamics, charge transfer should not be included within the classical polarization contribution and should preferably be either incorporated into the pairwise van der Waals contribution or treated explicitly.

I. Introduction

Since the 1940s, we have begun to appreciate that specific biological functions critically depend on the presence of

zinc.¹ Moreover, its divalent cation, Zn^{2+} , plays an important role in many metalloenzymes by acting directly as a structural element in proteins such as Zn-fingers² or by serving as a cofactor.³ Due to zinc’s soft character and the subtle nature of its interactions with the biological environment,⁴ quantum mechanics (QM) is usually the primary methodology for the study of Zn^{2+} –metalloproteins.^{5–7} Of course, such an approach is limited to “static” structures of relatively small biomimetic models due to the high computational demands

* Corresponding authors. E-mail: jpp@lct.jussieu.fr (J.-P.P.), pren@mail.utexas.edu (P.R.).

[†] University of Texas at Austin.

[‡] UPMC Univ. Paris.

[§] CNRS.

by state of the art QM approaches. Hybrid methods that combine QM and molecular mechanics (QM/MM)^{8–11} offer the possibility to treat the whole protein on longer time scales. Nevertheless, if one is interested in the dynamical behavior of Zn²⁺ complexes, available methods remain sparse. Traditional fixed charge force fields are unable to capture the interactions between Zn²⁺ and its ligands, or even to keep the Zn²⁺ “in place”, unless using artificial bonds¹² or extra charge sites.¹³ Recent studies based on quasi-chemical theory have shown the importance of polarization in ion hydration.^{14,15}

As an alternative to QM, anisotropic polarizable molecular mechanics (APMM) methods, such as SIBFA (sum of interactions between fragments *ab initio* computed)^{16,17} and AMOEBA,¹⁸ have been developed in recent years. Such techniques are computationally more efficient and provide potential energy surfaces in close agreement with QM. For the specific case of Zn²⁺, SIBFA, which treats both polarization and charge transfer contributions, has been shown to be particularly accurate and has enabled the study of large biological systems.^{16,19–23} As SIBFA’s extension to MD is under development, AMOEBA has already been extensively tested in simulations of various systems including proteins^{24–26} and has been shown to be particularly suited for the computation of dynamical properties of metal cations of biological interest.^{19–22,27–29}

In this contribution, as a first step toward modeling Zn²⁺ metalloenzymes, we will show that AMOEBA is able to accurately capture Zn²⁺ solvation properties. In the first part of this work, we will detail the parametrization process which is grounded on gas phase *ab initio* calculations following a “bottom-up” approach.¹⁶ The application of energy decomposition analyses (EDA) techniques¹⁷ such as the constrained space orbital variations (CSOV),³⁰ reduced variational space (RVS),³¹ and symmetry adapted perturbation theory (SAPT)³² to AMOEBA’s parametrization will be discussed. Moreover, such approaches are used to evaluate the importance of the charge transfer contribution. The nature of the interaction of Zn²⁺ with water will be investigated using the electron localization function (ELF)³³ topological analysis.³⁴ In the second part, we will perform extensive condensed-phase simulations using AMOEBA to compute Zn²⁺ solvation properties such as the ion–water radial distribution function (RDF), water residence times, and the coordination number, as well as the solvation free energy. Comparison is made to experimental results as well as other divalent cations that have previously been studied using AMOEBA.

II. Computational Details

Gas Phase *ab Initio* Calculations. The intermolecular interaction energies of Zn²⁺–H₂O at various separations were calculated using Gaussian 03³⁵ at the MP2(full) level. Basis set superposition error (BSSE) correction was included in the binding energy. The geometry of the previously derived AMOEBA water model was applied.^{36,37} The aug-cc-pVTZ basis set³⁸ was employed for water and the 6-31G(2d,2p) basis set for the Zn²⁺ cation. Post-Hartree–Fock symmetry adapted perturbation theory (SAPT) calculations were performed with the same basis sets at the MP2 and

CCSD levels using the *Dalton* package³⁸ and SAPT 96.³⁹ CSOV polarization energy calculations were performed using a modified version⁴⁰ of HONDO95.3⁴⁰ with the B3LYP methods^{41,42} using the above basis sets. The Zn²⁺ atomic polarizability was computed using Gaussian 03 at the MP2(full)/6-31G** level.

Additional energy decomposition analysis was performed on the zinc hydrated cluster with the reduced variational space (RVS) scheme as implemented in the GAMESS⁴³ software. The RVS energy decomposition computations were performed at the Hartree–Fock (HF) level using the CEP 4–31G(2d) basis set⁴⁴ augmented with two diffuse 3d polarization functions on heavy atoms (double- ζ -quality pseudopotential) and at the aug-cc-pVTZ basis set level (6-31G** for Zn(II)).

Electron Localization Function Analysis (ELF). In the framework of the ELF^{33,45} topological analysis,³⁴ the molecular space is divided into a set of molecular volumes or regions (the so-called “basins”) localized around maxima (attractors) of the vector field of the scalar ELF function. The ELF function can be interpreted as a signature of the electronic-pair distribution, and ELF is defined to have values restricted between 0 and 1 to facilitate its computation on a 3D grid and its interpretation. The core regions can be determined (if $Z > 2$) for any atom A. Regions associated to lone pairs are referred to as V(A), and bonding regions denoting chemical bonds are denoted V(A,B). The approach offers an evaluation of the basin electronic population as well as an evaluation of local electrostatic moments. It is also important to point out that metal cations exhibit a specific topological signature in the electron localization of their density interacting with ligands according to their “soft” or “hard” character. Indeed, a metal cation can split its outer-shell density (the so-called subvalent domains or basins) according to its capability to form a partly covalent bond involving charge transfer.⁴⁶ More details about the ELF function and its application to biology can be found in a recent review.⁴⁷ All computations have been performed using a modified version⁴⁸ of the Top-Mod package.⁴⁹

III. Parameterization and Free Energy Simulations

Use of CSOV and SAPT Energy Decompositions Schemes. Following a procedure that has already shown success with Ca(II) and Mg(II),²⁹ the Zn²⁺ cation is parametrized by first matching the distance dependence of AMOEBA polarization energies of the ion–water dimer in the gas phase with reference *ab initio* CSOV polarization energy results. In order to supplement the CSOV decomposition, we have also performed SAPT computations (available in the Supporting Information). It is important to note that, despite the fact that SAPT could be expected to be the reference analysis offering up to CCSD correlation corrections to compute the contributions, a close examination of the results clearly shows that SAPT has problems with converging to the supermolecular interaction energy. A similar trend has recently been observed by Rayon et al.⁶ It appears that difficulty with convergence is mainly due to the second order induction term, which consists of both

polarization and charge transfer energies.¹⁷ Such a problem is not new, as Claverie⁵⁰ and then Kutzelnigg⁵¹ showed 30 years ago that the convergence of the SAPT expansion was not guaranteed. As in recent studies on water,⁵² the discrepancy of total SAPT energies compared to supermolecular interaction energy results can be traced back to the importance of the third order induction correction. Their inclusion clearly enhances the binding energy and could therefore improve SAPT results. We reported here extensive SAPT results in a detailed Supporting Information section dealing with the Zn²⁺–water complex. As one can see from the Supporting Information, at the Hartree–Fock level, the SAPT approximation tends not to converge at short-range, the total SAPT energy being far from the supermolecular HF value. Around the equilibrium (Zn²⁺–O = 2.0 Å), and beyond, this discrepancy tends to diminish, becoming negligible at long range. However, since the AMOEBA force field is based on the reproduction of supermolecular interaction energies, we need short-range induction data in order to refine the parameters. Moreover, as SAPT induction embodies both charge transfer and polarization, we cannot fit directly the sole “polarization only” Thole model to these values. Consequently, for the present purpose of AMOEBA’s fitting, we have limited our use of the SAPT results to a comparison of the accuracy of AMOEBA’s Halgren 14–7 van der Waals function⁵³ at long-range directly to the sum of the SAPT exchange-repulsion, dispersion, and exchange-dispersion. Such a fit is reflected in the good agreement between the AMOEBA and *ab initio* total interaction energy at long range (see Figure 5).

In summary, we fit AMOEBA’s polarization contribution (the damping factor “*a*” in the next section) to the CSOV results. The remaining induction contribution (charge transfer) will be included in the van der Waals term as a result of matching the total binding energy of AMOEBA to that of QM. In the absence of an explicit charge transfer term, such a strategy is justified, as the charge transfer contribution is notably smaller in magnitude compared to polarization^{4,16,22,23} and a good percentage of it (namely the two-body part) could be accurately included within AMOEBA’s van der Waals term assuming that many-body charge transfer is not the driving force of Zn(II) solvation dynamics. The validity of such an assumption and the applicability of the present parametrization scheme to Zn²⁺ will be discussed in the first section of the discussion.

AMOEBA Calculation Details. The AMOEBA polarizable force field^{28,36,37} is used to study the solvation dynamics of Zn(II). Hence, the electrostatic term of the model accounts for polarizability via atomic dipole induction:

$$\mu_{i,\alpha}^{\text{ind}} = \alpha_i \left(\sum_{\{j\}} T_{\alpha}^{ij} M_j + \sum_{\{j'\}} T_{\alpha\beta}^{ij'} \mu_{j',\beta}^{\text{ind}} \right) \text{ for } \alpha, \beta = 1, 2, 3$$

where $M_j = [q_j, \mu_{j,1}, \mu_{j,2}, \mu_{j,3}, \dots]^T$ are the permanent charge, dipole, and quadrupole moments and $T_{\alpha}^{ij} = [T_{\alpha}, T_{\alpha 1}, T_{\alpha 2}, T_{\alpha 3}, \dots]$ is the interaction matrix between atoms *i* and *j*. The Einstein convention is used to sum over indices α and β . The atomic polarizability, α_i , is parametrized for the zinc cation in this work. Note that the first term within the parentheses corresponds to the polarization field due to permanent

multipoles, while the second term corresponds to the polarization field due to induced dipoles produced at the other atoms.

The dipole polarization is damped via smeared charge distributions as proposed by Thole:⁵⁴

$$\rho = \frac{3a}{4\pi} \exp(-au^3)$$

where $u = R_{ij}/(\alpha_i \alpha_j)^{1/6}$ is the effective distance between atoms *i* and *j*. The scalar *a*, a dimensionless parameter corresponding to the width of the smeared charge distribution, is parametrized to be 0.39 for water³⁶ and monovalent ions.⁵⁵ A previous study suggested that, for monovalent ions, AMOEBA is able to reproduce *ab initio* MP2 correlated results and hydration enthalpies without modifying the damping factor. However, since divalent ions, such as Ca²⁺ and Mg²⁺,^{28,29} require a wider charge distribution in order to agree with QM ion–water dimer energy, smaller values of *a* were assigned. The value for Zn²⁺ is also adjusted from 0.39 and is compared with those of Ca²⁺ and Mg²⁺ below.

The repulsion–dispersion (van der Waals) interaction is represented by a buffered 14–7 function:⁵³

$$U_{ij}^{\text{buff}} = \varepsilon_{ij} \left(\frac{1 + \delta}{\rho_{ij} + \delta} \right)^{n-m} \left(\frac{1 + \gamma}{\rho_{ij}^m + \gamma} - 2 \right)$$

where ε_{ij} is the potential well depth. In addition, ρ_{ij} is R_{ij}/R_{ij}^0 , where R_{ij} is the separation distance between atoms *i* and *j*, and R_{ij}^0 is the minimum energy distance. Following Halgren, we used fixed values of $n = 14$, $m = 7$, $\delta = 0.07$, and $\gamma = 0.12$. The values for R_{ij}^0 and ε_{ij} are parametrized. The polarizable water model as developed by Ren and Ponder³⁶ is employed in this study.

With water geometry fixed, the Zn²⁺–O distances were varied between 1.5 and 5 Å. The damping factor “*a*” was adjusted so that the AMOEBA polarization energy matched the CSOV values as much as possible. Next, parameters for the van der Waals interaction, R^0 (radius) and ε (well-depth), were derived by comparing the total ion–water binding energy computed by AMOEBA to the *ab initio* values at various distances. For interactions between different types of atoms, these parameters undergo combination rules as described by Ponder et al.²⁶ The binding energies were computed as the total energy less the isolated water and ion energies at an infinite separation distance.

Molecular dynamics simulations were performed via the TINKER 5 package⁵⁶ to compute the solvation free energy of Zn²⁺. Fourteen independent simulations were first performed to “grow” the Zn vdW particle by gradually varying $R(\lambda) = \lambda(R_{\text{final}})$ and $\varepsilon(\lambda) = \lambda(\varepsilon_{\text{final}})$, where $\lambda = (0.0, 0.0001, 0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)$. Subsequently, 30 simulations were performed to “grow” the (+2) charge of Zn²⁺ along with its polarizability such that $q(\lambda') = \lambda'(q_{\text{final}})$ and $\alpha(\lambda') = \lambda'(\alpha_{\text{final}})$, where $\lambda' = (0.0, 0.1, 0.2, 0.3, 0.325, 0.350, 0.375, 0.400, 0.425, 0.450, 0.475, 0.500, \dots, 1.0)$. The long-range electrostatics are modeled with particle-mesh Ewald summation for atomic multipoles with a cutoff of 7 Å in real space and 0.5 Å spacing and a fifth-order spline in reciprocal space.⁵⁷ The convergence

Table 1. Polarization Energy and Charge Transfer Energy from Restricted Variational Space (RVS) Energy Decomposition of Zn^{2+} in the Presence of Water Clusters of Sizes 1, 4, 5, and 6 at the HF/CEP-41G(2d) Level (or HF/aug-cc-PVTZ/6-31G**, Results in Parentheses)^a

complex	$\text{Zn}(\text{H}_2\text{O})$	$[\text{Zn}(\text{H}_2\text{O})_4]^{2+}$	$[\text{Zn}(\text{H}_2\text{O})_5]^{2+}$	$[\text{Zn}(\text{H}_2\text{O})_6]^{2+}$
E_{pol} (RVS)	-37.6	-118.7 (-135.3)	-110.8 (-127.5)	-104.3 (-117.5)
E_{CT} (RVS)	-10.9	-28.7 (-9.3)	-24.5 (-6.7)	-21.8 (-4.51)
$(E_{\text{CT}}/(E_{\text{pol}} + E_{\text{CT}})) \times 100$	22.5	19.4 (6.4)	18.1 (5.0)	16.6 (3.7)

^a Percentage of induction energy due to charge transfer is presented in the last row. All are in units of kcal/mol.

criteria for induced dipole computation is 0.01 D. Molecular dynamics simulations were performed with a 1 fs time step for 500 ps at each perturbation step. Trajectories were saved every 0.1 ps after the first 50 ps equilibration period. The temperature was maintained using the Berendsen weak coupling method at 298 K.⁵⁸ The system contained 512 water molecules with one Zn^{2+} ion, and 24.857 Å is the length of each side of the cube.

The absolute free energy was computed from the perturbation steps by using the Bennett acceptance ratio (BAR), a free energy calculation method that utilizes forward and reverse perturbations to minimize variance.^{59,60} MD simulations were extended for 2.2 ns (total 2.7 ns) with the final Zn^{2+} parameters, and the resulting trajectory was used in the analysis of the structure and dynamics of water molecules in the first solvation shell. Water molecules separated by a distance less than the first minimum of the $\text{Zn}^{2+}-\text{O}$ RDF were considered to be in the first solvation shell. The averaged residence time of the first shell water molecules was directly measured by monitoring the entering and exiting events.

IV. Results and Discussion

Contribution of Charge Transfer in Zn^{2+} -Water Complexes. The lack of explicit charge transfer (CT) in AMOEBA presents an interesting challenge. When the CT contribution is significant, despite its limited magnitude in many-body complexes, it may be difficult to capture the overall many-body effect by only considering polarization. Therefore, it is important to investigate the CT contribution to the Zn^{2+} -water interaction energy and its dependence on the system size. To estimate the magnitude of charge transfer, we performed several RVS energy decomposition analyses on complexes up to $[\text{Zn}(\text{H}_2\text{O})_6]^{2+}$.

We report here complexes that were initially studied by Gresh et al.^{22,61,62} the monoligated $[\text{Zn}(\text{H}_2\text{O})]^{2+}$ complex and polyligated $[\text{Zn}(\text{H}_2\text{O})_6]^{2+}$, $[\text{Zn}(\text{H}_2\text{O})_5(\text{H}_2\text{O})]^{2+}$, and $[\text{Zn}(\text{H}_2\text{O})_4(\text{H}_2\text{O})_2]^{2+}$ arrangements (octahedral \rightarrow pyramidal \rightarrow tetrahedral first-shell). As we can see in Table 1, the importance of charge transfer relative to polarization varies with the size of the $\text{Zn}^{2+}-\text{(H}_2\text{O)}_n$ complex and depends on the basis set. It makes up a significant portion of induction for a monoligated $[\text{Zn}(\text{H}_2\text{O})]^{2+}$, and its contribution decreases as the number of ligating water molecules increases to 6. The charge transfer effect appears to be diluted within the entire induction energy (polarization and charge transfer) as the number of water molecules grows in agreement with the previous observation of anticooperative effects.^{22,61,62} Note that basis set superposition error (BSSE) is not taken into account. As indicated by Stone,⁶³ such systematic error can

be clearly associated with the charge transfer effect. In contrast to the inverse relationship between CT and water ligation expressed by the zinc cation, the CT contribution associated with anions, such as Cl^- , has been observed to increase as ligation increases.⁶⁴ This phenomenon may be due to the asymmetric solvation environment for the anions as well as their modes of water ligation. However, analyses of CT effects are not apparent, as they are found in both induction energy and basis set superposition error.⁶³ For the largest complex $[\text{Zn}(\text{H}_2\text{O})_6]^{2+}$, the BSSE amounts to 3.3 kcal/mol. If removed, the relative weight of charge transfer to the total induction reduces from 16.6% (Table 1) to 15.3% at the CEP-31G(2d) level. Using the large aug-cc-PVTZ for water coupled to the 6-31G** basis set for $\text{Zn}(\text{II})$, the observed trends are even more pronounced as the relative importance of charge transfer strongly diminishes from 6.4% of the whole induction for $[\text{Zn}(\text{H}_2\text{O})_4]^{2+}$ to less than 4% for the $[\text{Zn}(\text{H}_2\text{O})_6]^{2+}$ complex while polarization becomes more dominant. Thus, the magnitude of the CT estimated by *ab initio* methods is greatly dependent on the basis set used. While our results have been obtained at the Hartree-Fock level, recent studies clearly show that correlation acts on induction and leads to greater charge transfer energy.^{17,40} For this reason, we computed the induction energies on selected water clusters at both the HF and DFT level using a recently introduced energy decomposition analysis (EDA) technique based on single configuration-interaction (CI) localized fragment orbitals.⁶⁵ We indeed find that the CT contribution increases slightly with DFT; however, overall it accounts for less than 20% of the total induction energy for monoligated complexes and presumably would be even less in the bulk water environment.

To gain further insight into the interaction of Zn^{2+} with water, we performed the electron localization function (ELF) analysis. An important asset of the ELF topological analysis is that it provides a clear description of a covalent bond between two atoms as it exhibits a basin between atoms to indicate electron sharing. Here, we have considered several $\text{Zn}^{2+}-\text{(water)}_n$ complexes, $n = 1-6$. An important discovery from ELF analysis is that a covalent $\text{V}(\text{Zn},\text{O})$ is only observed in the monoligated Zn^{2+} -water complex (Figure 1). In that case, we observe a net concentration of electrons between the zinc cation and the water oxygen, a clear sign of covalent bonding ($1.9 e^-$ on the bond). As n increases, the covalent $\text{V}(\text{Zn},\text{O})$ feature disappears despite a residual mixing of Zn^{2+} contributions in the oxygen basin. Indeed, as the $\text{Zn}-\text{O}$ distances increase with n (Figures 2 and 3), the $\text{Zn}-\text{O}$ bond becomes more ionic as the charge transfer quickly diminishes. Such behavior could be then understood using the subvalence concept.⁴⁶ As shown by de Courcy

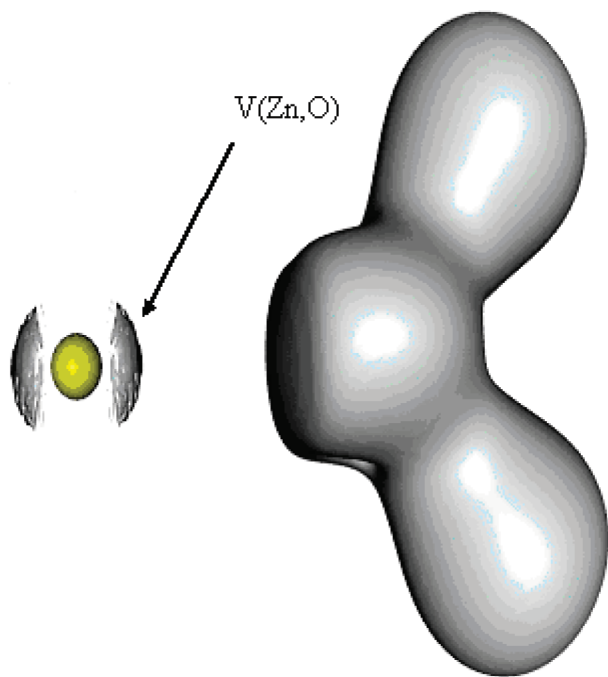


Figure 1. ELF localization domains (basins) for the $\text{Zn}^{2+}-\text{H}_2\text{O}$ complex. A covalent $V(\text{Zn},\text{O})$ basin reflecting electron sharing is observed and reveals the covalent nature of the Zn–O interaction.

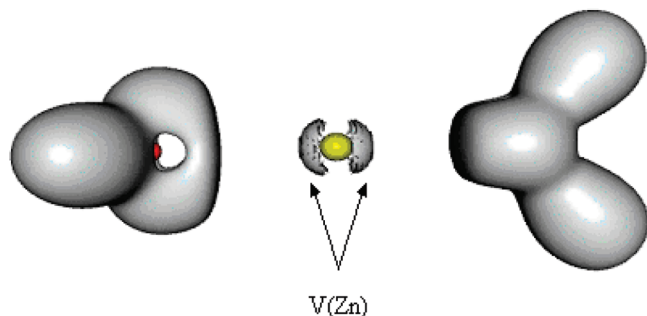


Figure 2. ELF localization domains (basins) for the $\text{Zn}^{2+}-\text{(H}_2\text{O)}_2$ complex. Noncovalent $V(\text{Zn})$ basins are observed describing the deformation of Zn^{2+} outer-shells' density within the fields of the water molecules.

et al.,^{4,46} the cation density is split into several “subvalent” domains as its outer shells appear strongly polarized, which explains why covalency is not achieved. If the cation electron density is strongly delocalized toward the oxygen atoms, the center of the basin remains closer to Zn^{2+} (covalent bonding would implicate a polarized bond with a covalent $V(\text{Zn},\text{O})$ basin localized closer to the more electronegative oxygen). ELF results thus suggest that, although the induction in the Zn^{2+} –water monoligated complex is dominated by charge transfer, this is not to the case for n from 2 to 6. In the latter case, the many-body effects are driven by the Zn^{2+} outer shells' plasticity that accommodates the strongly polarized water molecules. The atoms in molecules (AIM) population analysis confirms that such behavior is present in DFT as well as at the MP2 level. As expected (see refs 6 and 40, for example), DFT tends to slightly overbind the complexes as compared to MP2, which clearly gives a better description of the bonding over Hartree–Fock.

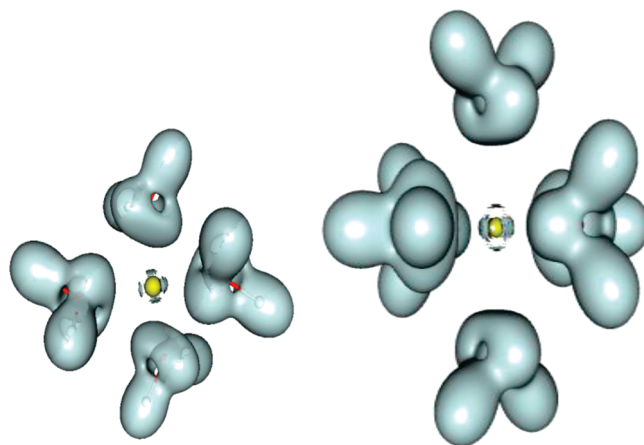


Figure 3. ELF localization domains (basins) for the $\text{Zn}^{2+}-\text{(H}_2\text{O)}_4$ and $\text{Zn}^{2+}-\text{(H}_2\text{O)}_6$ complexes. Again, noncovalent $V(\text{Zn})$ basins are observed.

Table 2. Ion Parameters: Diameter, Well Depth, Polarizability, and Dimensionless Damping Coefficient

ion	R (Å)	ϵ (kcal/mol)	α (Å ³)	a^a
Zn^{2+}	2.68	0.222	0.260	0.2096
Mg^{2+}	2.94	0.300	0.080	0.0952
Ca^{2+}	3.63	0.350	0.550	0.1585

^a a is the dimensionless damping coefficient.

To conclude on these various results, we expect that AMOEBA will improve in accuracy with an increase in system size as the charge transfer effect becomes less important and the total induction will be dominated by polarization. In other words, we anticipate the discrepancy between AMOEBA and QM observed in the monoligated water– Zn^{2+} complex to disappear in the condensed phase. This also suggests that an “ad-hoc” inclusion of the charge transfer into the polarization contribution by adjusting the polarization damping scheme (see the Thole model in the Computational Details) is probably not a suitable strategy. Indeed, charge transfer can rapidly vanish, and “polarization only” models overfitted on monoligated complexes to include charge transfer will lead to an overestimated many-body effect in bulk-phase simulation as the polarization would still contain the unphysical charge transfer. Charge transfer should be treated explicitly or included in the van der Waals to a certain extent. In this study, we adopt the latter approach to effectively incorporate the charge transfer in the bulk environment into the vdW interactions.

Accuracy of the AMOEBA Parametrization. The distance-dependent dimer binding energies were used to adjust vdW parameters (R and ϵ), and the damping factor of polarizability (a) for Zn^{2+} was adjusted to match the CSOV polarization energy. Table 2 lists the final parameters of the Zn^{2+} cation as well as the Mg^{2+} and Ca^{2+} cations parametrized by Jiao et al.²⁸ that are optimized for the Tinker implementation of AMOEBA. Meanwhile, parameters optimized for a slightly modified implementation of the AMOEBA force field present in Amber which embodies a modified periodic boundary condition treatment of long-range van der Waals are available as well.²⁹ It should be noted that, although the previously reported parameters for Mg^{2+}

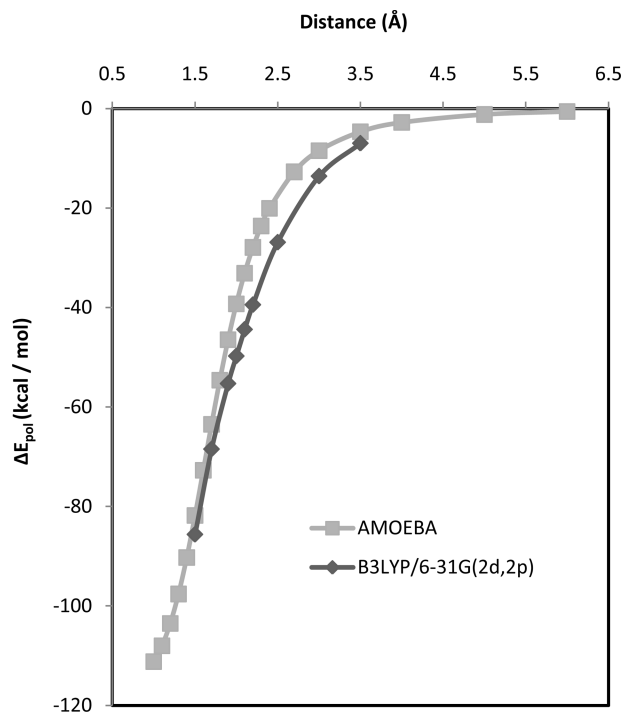


Figure 4. Polarization energy of the zinc and water dimer in the gas phase as a function of separation distance.

and Ca^{2+} contained typographic inconsistencies,²⁸ results from that work (thermodynamic energy, structural analysis, etc.) are obtained from parameters consistent with Table 2. Figure 4 compares CSOV polarization energy calculations with the AMOEBA polarizable force field as a function of distance between the cation and water. The difference between the two methods is mainly found at distances between 2 and 3 Å, where the charge transfer effect in the two-body system is strong. However, such discrepancy is expected to diminish in bulk water as the charge transfer effect is expected to be less important, as explained above. Comparisons between total binding energies of the AMOEBA polarizable model and *ab initio* calculations are shown in Figure 5. As expected, the interaction energy between 2 and 3.5 Å appears to be underestimated (less negative) compared to the *ab initio* result. The strategy here is, however, not to overfit the AMOEBA model to the monoligated Zn^{2+} complex, as the polarization energy and total interaction energy are already very reasonable considering the relatively simple force field functional form. The AMOEBA association energies for $[\text{Zn}(\text{H}_2\text{O})_6]^{2+}$, $[\text{Zn}(\text{H}_2\text{O})_5(\text{H}_2\text{O})]^{2+}$, and $[\text{Zn}(\text{H}_2\text{O})_4(\text{H}_2\text{O})_2]^{2+}$ complexes are -334.4 , -333.4 , and $-331.9/-333.7$ kcal/mol, respectively. Given that AMOEBA is mainly targeting the condensed phase, the trend observed here is in reasonable agreement with the previous *ab initio* results (-345.3 , -341.3 , $-337.4/-337.8$ kcal/mol using CEP 4-31G (2d) basis set; -365.9 , -363.3 , $-360.0/-362.4$ kcal/mol using 6-311G** basis set).⁶² Our approach is further validated in the condensed-phase hydration properties calculation next.

Evaluation of Zn^{2+} Solvation in Water Using AMOEBA. The hydration free energy is the key quantity describing the thermodynamic stability of an ion in solution. The solvation free energy of zinc in water has been computed

from molecular dynamics simulations using free energy perturbation (FEP). Table 3 lists the free energy of hydration for Zn^{2+} , Mg^{2+} , and Ca^{2+} compared with experiment-derived values^{66,67} and the results from the quasi-chemical approximation method.¹⁴ The free energy values computed from AMOEBA are closer to those from quasi-chemical approximation (QCA) than to the data interpreted from experimental measurement. In the QCA method, the region around the solute of interest is partitioned into inner and outer shell domains. The inner shell was treated quantum mechanically, while the outer shell was evaluated using a dielectric continuum model. Note that, to decompose the hydration free energy of a neutral ion pair, tetraphenylarsonium tetraphenylborate (TATB) has been most widely chosen as a reference salt, on the basis of the extra thermodynamic assumption that the large and hydrophobic ions do not produce charge-specific solvent ordering effects.^{55,66} Our results show better agreement with “experimental values” for Ca^{2+} and Mg^{2+} ions by Schmid et al., who derived the single ion hydration free energy by using the theoretically determined proton hydration free energy as a reference.⁶⁷ The hydration free energy for the Zn^{2+} ion computed using AMOEBA is in good agreement with values given by Marcus⁶⁶ and Asthagiri et al.,¹⁴ with deviations less than 1.9% and 0.2%, respectively.

Solvent Structure and Dynamics. To characterize the structure of water molecules around the ion, the radial distribution function (RDF) between the Zn^{2+} and oxygen atom of the water molecule has been obtained from the 2.7 ns molecular dynamics simulation (Figure 6). The running integration of Zn–O, which imparts water–ion coordination information, is also plotted. The first minimum in the ion–O RDF is at a distance of 2.85 Å, which can be interpreted as the effective “size” of the complex composed of the ion and first water solvent shell. The running integration indicates a water-coordination number of 6 in the first solvation shell, which is consistent with experimental observations.^{68–73} As expected, the zinc cation binds to the first water shell more tightly than other ions, as evident in the more pronounced and narrow first peak as well as the shortest separation, as shown in the ion–O RDFs in Figure 7. Overall, the zinc solvation structure show greater similarity to Mg^{2+} than Ca^{2+} .

The Born theory of ion solvation⁷⁴ states that there exists an effective solvation radius, R_B , for each ion such that the solvation free energy of the ion in a dielectric medium is given by

$$\Delta A = -\frac{q^2}{2R_B} \left(1 - \frac{1}{\epsilon_d} \right)$$

where q is the charge of the ion and ϵ_d is the dielectric constant of the medium (80 for water). We have calculated the effective radius of zinc on the basis of the Born equation from the solvation free energy obtained from our simulations. Table 4 gives a detailed comparison among Zn^{2+} , Mg^{2+} , and Ca^{2+} . It should be noted, however, that previous studies have shown that ion hydration energy is not symmetric with respect to electronegativity,^{27,75,76} as is implied by the Born theory. The first peak of the Zn^{2+} –O RDF is at 1.98 Å, and

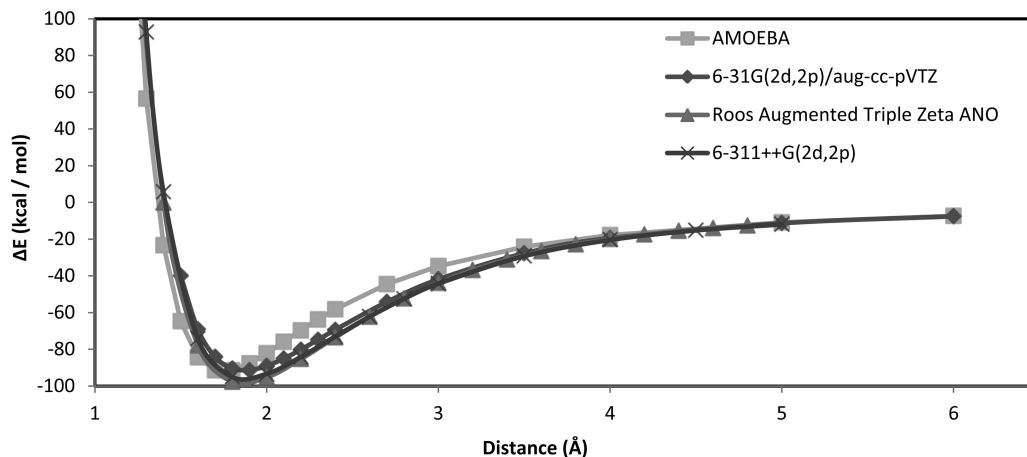


Figure 5. Binding energy of the zinc and water dimer in the gas phase as a function of separation distance. The 6-31G(2d,2p)/aug-cc-pVTZ indicates that 6-31G(2d,2p) was used to represent the Zn^{2+} cation and aug-cc-pVTZ was used to represent the water molecule. Binding energy obtained from the last two basis sets used the same basis sets for both the ion and water.

Table 3. Solvation Free Energy of Zinc in Water^a

ion	ΔG (kcal/mol)	experimental	quasi-chemical ^b
Zn^{2+}	-458.9(4.4)	-467.7 ^c	-460.0
Mg^{2+}	-431.1(2.9)	-435.4 ^d	-435.2
Ca^{2+}	-354.9(1.7)	-357.2 ^d	-356.6

^a A 1 mol L⁻¹ solution is chosen as the standard state. ^b Ref 14. ^c Ref 66. ^d Ref 67.

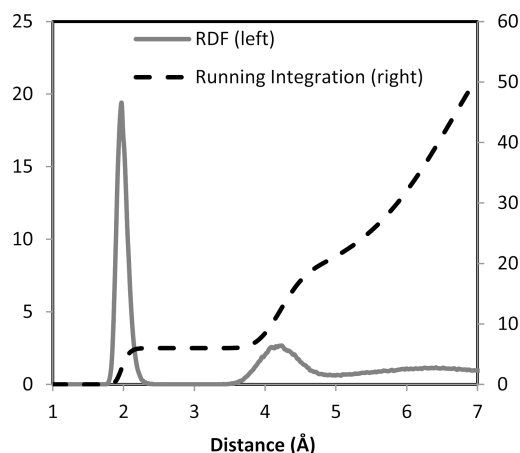


Figure 6. Radial distribution function of Zn^{2+} -O (left axis) and water coordination number (right axis).

the effective Born radius of the cation is calculated to be 1.47 Å. A difference of ~ 0.5 Å between the two quantities is consistent with the results of other mono- and divalent metal ions.^{27,28,77-79} The difference between the first minimum in the Zn^{2+} -O RDF and the Born radius is 1.38 Å and is consistent with studies of other ions as well.^{27,28}

In addition to the RDF, the solvation structure has been analyzed from the distribution of the angles formed by O-ion-O in the first water shell. Figure 8 compares the distribution of angles for Zn^{2+} , Mg^{2+} , and Ca^{2+} cations. With sharp peaks located near 90° and 180°, the distribution of the O- Zn^{2+} -O angle suggests a rigid octahedron geometry with the Zn^{2+} surrounded by six water molecules. Mg^{2+} shares a similar but slightly more flexible geometry, while results for Ca^{2+} suggest a more amorphous structure. Figure 9 shows the dipole moment at each distance (Å)

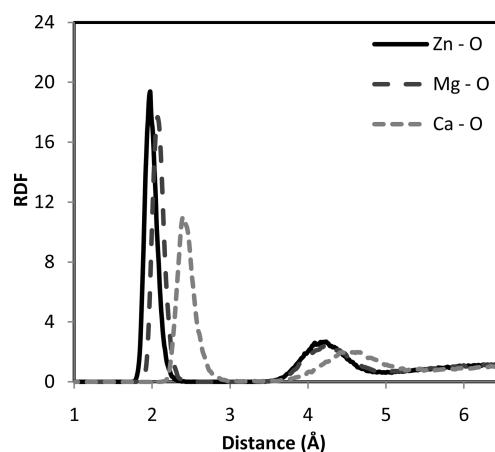


Figure 7. Radial distribution function of divalent cations (Zn^{2+} , Mg^{2+} , and Ca^{2+}) and the oxygen atom in water.

around the ion. Figure 10 is a sample frame from the molecular dynamics simulation to illustrate the octahedron arrangement between the zinc and the first shell water molecules.

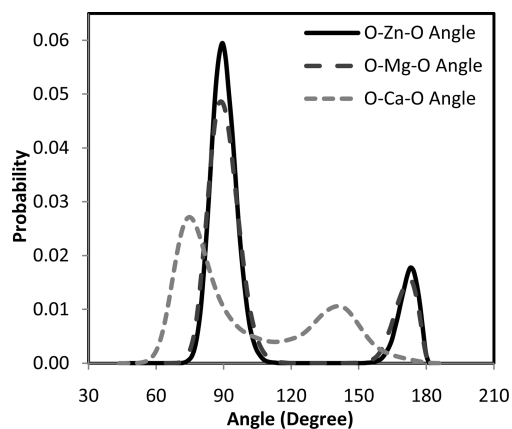
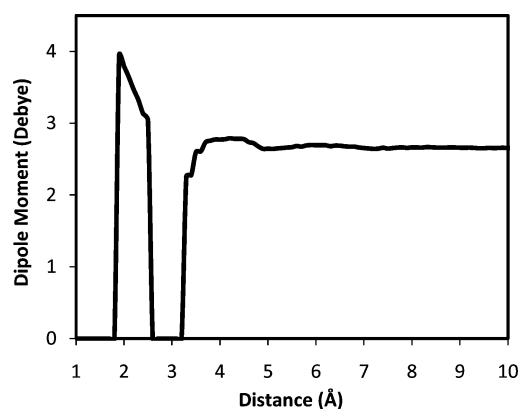
Dipole Moment. The average dipole moment of water as a function of distance away from the zinc cation is computed. At the closest distance of 1.9–2.5 Å, water experiences a dipole moment from 3.0 to 3.9 D. Due to the highly organized structure of the first water shell, a “vacuum” space free of water molecules is observed between 2.6 and 3.2 Å away from the cation, also evident in the Zn^{2+} -O RDF. The higher dipole moment of Zn^{2+} relative to bulk water (2.77 D³⁶) within the first water shell is consistent with a previous observation of other divalent cations.²⁸ The dipole moment of water in the first solvation shell of monovalent cations such as K^+ and Na^+ , however, is lower than that of bulk water.⁵⁵

Residence Time. We have investigated the lifetime of ion-water coordination by directly examining the average amount of time that a water molecule resides within the first solvation shell. The first solvation shell is determined by the position of the first minimum of the Zn-O RDF. If an oxygen atom is less than 2.85 Å away from the Zn^{2+} , the water is considered to be in the first solvation shell. Cutoff

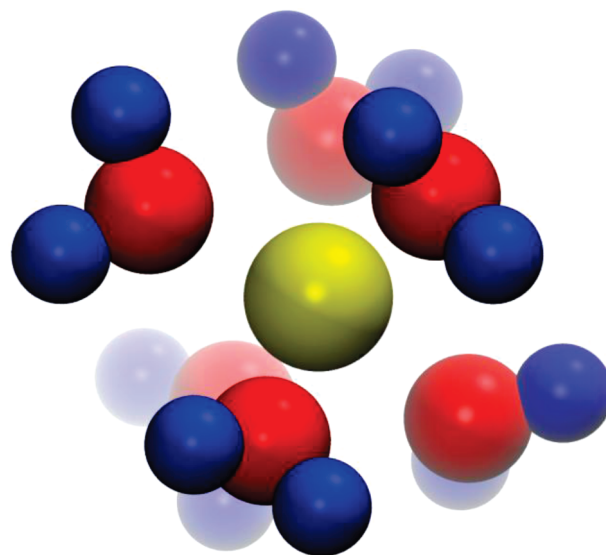
Table 4. Radii Results for Zn^{2+} , Mg^{2+} , and Ca^{2+} Cations^a

ion	Born radius (Å)	first peak in ion-O RDF	experimental first peak in ion-O RDF	QM/MM first peak	first minimum in ion-O RDF
Zn^{2+}	1.47	1.98	2.07 ^b	2.11–2.18 ^b	2.85
Mg^{2+}	1.56 ^c	2.07	2.09 ^d	2.13 ^e	2.95
Ca^{2+}	1.89 ^c	2.41	2.41–2.44;2.437;2.46 ^d	2.43–2.44 ^e	3.23

^a Born radii, first peak in ion-O RDF with AMOEBA polarizable force field, experimental first peak in ion-O RDF, and first minimum in ion-O RDF are all indicated in Å. ^b Ref 69. ^c Ref 28 ^d Refs 77, 78, and 79. ^e Refs 77, 78.

**Figure 8.** Water–ion–water angle distribution of divalent cations (Zn^{2+} , Mg^{2+} , and Ca^{2+}) and the oxygen atom in water.**Figure 9.** Dipole moment at each distance (Å) around the ion.

distances used for the first solvation shells of Mg^{2+} and Ca^{2+} are 2.95 and 3.23 Å, respectively. In Table 5, coordination numbers and residence times from AMOEBA simulations are compared with experimental values for Zn^{2+} , Mg^{2+} , and Ca^{2+} .^{20,80–86} The Zn^{2+} to water-proton dynamics are studied with quasi-elastic neutron scattering methods (QENS) as described by Salmon et al.⁸⁰ The water residence times directly sampled from the MD simulations are in better agreement with experimental results than those previously inferred from the time correlation function of the instantaneous first shell coordination number.²⁸ According to AMOEBA simulations, the residence time in the first solvation shell around Zn^{2+} is at least 2 ns, and the water molecules around Ca^{2+} have a lifetime on the order of several picoseconds, both of which are within the experimental ranges. For Mg^{2+} , experimental results suggest that water molecules could live up to a few microseconds, while the simulations using AMOEBA indicate a residence time similar

**Figure 10.** First solvation shell around the Zn^{2+} ion.**Table 5.** Coordination Number, Experimental Coordination Number, Residence Time, Experimental Residence Time, and QM/MM Residence Times for Each Type of Divalent Cations.

ion	coordination number	exp. coordination number	residence time (s)	exp residence time (s)
Zn^{2+}	6	6 ^a	2.2×10^{-9}	10^{-10} to 10^{-9b}
Mg^{2+}	6	6 ^c	1.9×10^{-9}	2×10^{-6} to $10^{-5d,e}$
Ca^{2+}	7.3	7.2 ± 1.2^f	1.33×10^{-10}	$<10^{-10}$ to 10^{-7e}

^a Refs 68–73. ^b Refs 80 and 20. ^c Ref 85. ^d Refs 81. ^e Refs 84, 82 and references within, and 83. ^f Ref 86.

to that of Zn^{2+} . Classical fixed-charge molecular mechanic methods suggest a residence time of 146 ps⁸⁷ for water around Zn^{2+} , while quantum mechanical methods have not attained simulation times long enough to observe the exchange of water molecules in the first shell.^{68,88} The calculated water residence times are consistent with the analyses of the radial distribution function and water angle distribution. A longer residence time is accompanied by a more ordered and closely packed water structure near the cation.

Conclusions

We showed in this contribution that AMOEBA was able to provide a reasonably accurate description of Zn^{2+} interaction with water, especially in the bulk water environment. We explained in detail one of the reasons for such good performance—the *ab initio* calculations demonstrated that the relative importance of charge transfer diminishes as the number of water molecule increases, a sign of anticooperativity. We have established a fitting strategy for induction:

charge transfer can be included in the pairwise dispersion in the van der Waals contribution; incorporation of charge transfer into polarization would lead to an overestimation of the many-body effects. Despite the difficulty involved with the AMOEBA model reproducing the binding energy of the monoligated Zn^{2+} -water complex, which exhibits nonclassical covalent bonding as shown by ELF topological analysis, AMOEBA is able to afford robust estimation of the hydration free energy along with reasonable solvation structure and dynamics. The current and previous studies suggest that the classical polarizable multipole-based AMOEBA is an effective tool for modeling ions in bulk solution, as good relative solvation free energies, structures, and dynamic properties have been obtained for a range of mono- and divalent cations. The work clearly demonstrates the need for “interpretative” *ab initio* techniques (ELF, EDA methods) in order to follow a bottom-up approach going from the gas-phase *ab initio* calculations to condensed-phase MD simulations. In addition, the zinc model developed in this work opens the door for future study of zinc-containing metalloproteins. Further investigation is necessary to determine whether the presence of negatively charged species interacting with Zn^{2+} would require an explicit consideration of charge transfer contribution in the classical energy function.

Acknowledgment. This research was supported by grants from the National Institute of General Medical Sciences (R01GM079686) and the Robert A. Welch Foundation (F-1691) to P.R. This work was also supported by the French National Research Agency (ANR) on project LASIHMODO (ANR-08-BLAN-0158-01) (J.-P.P.). Some computations have been carried out at GENCI IDRIS (F. 91403 Orsay, France) and CRIHAN (F. 76800 Saint-Etienne-de-Rouvray, France) supercomputer centers.

Supporting Information Available: Symmetry adapted perturbation theory (SAPT) results for the water-Zn(II) complex. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Keilin, D.; Mann, T. Carbonic anhydrase. Purification and nature of the enzyme. *Biochem. J.* **1940**, *34*, 1163–1176.
- (2) Maynard, A. T.; Covell, D. G. Reactivity of zinc finger cores: Analysis of protein packing and electrostatic screening. *J. Am. Chem. Soc.* **2001**, *123*, 1047–1058.
- (3) Lipscomb, W. N.; Strater, N. Recent advances in zinc enzymology. *Chem. Rev. (Washington, DC, U. S.)* **1996**, *96*, 2375–2433.
- (4) De Courcy, B.; Gresh, N.; Piquemal, J. P. Importance of lone pair interactions/redistribution in hard and soft ligands within the active site of alcohol dehydrogenase Zn-metalloenzyme: Insights from electron localization function. *Interdisc. Sci.: Comput. Life Sci.* **2009**, *1*, 55–60.
- (5) Gresh, N.; Garmer, D. R. Comparative binding energetics of Mg^{2+} , Ca^{2+} , Zn^{2+} , and Cd^{2+} to biologically relevant ligands: Combined *ab initio* SCF supermolecule and molecular mechanics investigation. *J. Comput. Chem.* **1996**, *17*, 1481–1495.
- (6) Rayon, V. M.; Valdes, H.; Diaz, N.; Suarez, D. Monoligated Zn(II) complexes: *Ab initio* benchmark calculations and comparison with density functional theory methodologies. *J. Chem. Theory Comput.* **2008**, *4*, 243–256.
- (7) Amin, E. A.; Truhlar, D. G. Zn coordination chemistry: Development of benchmark suites for geometries, dipole moments, and bond dissociation energies and their use to test and validate density functionals and molecular orbital theory. *J. Chem. Theory Comput.* **2008**, *4*, 75–85.
- (8) Warshel, A.; Levitt, M. Theoretical studies of enzymic reactions - dielectric, electrostatic and steric stabilization of carbonium-ion in reaction of lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–249.
- (9) Estiu, G.; Suarez, D.; Merz, K. M. Quantum mechanical and molecular dynamics simulations of ureases and Zn beta-lactamases. *J. Comput. Chem.* **2006**, *27*, 1240–1262.
- (10) Friesner, R. A.; Guallar, V. *Ab initio* quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.
- (11) Ryde, U. Combined quantum and molecular mechanics calculations on metalloproteins. *Curr. Opin. Chem. Biol.* **2003**, *7*, 136–142.
- (12) Tuccinardi, T.; Martinelli, A.; Nuti, E.; Carelli, P.; Balzano, F.; Uccello-Barretta, G.; Murphy, G.; Rossello, A. Amber force field implementation, molecular modelling study, synthesis and MMP-1/MMP-2 inhibition profile of (R) and (S)-N-hydroxy-2-(N-isopropoxybiphenyl-4-ylsulfonamido)-3-methylbutanamides. *Bioorg. Med. Chem.* **2006**, *14*, 4260–4276.
- (13) Yuan-Ping, P. Successful molecular dynamics simulation of two zinc complexes bridged by a hydroxide in phosphotriesterase using the cationic dummy atom method. *Proteins: Struct., Funct., Genet.* **2001**, *45*, 183–189.
- (14) Asthagiri, D.; Pratt, L. R.; Paulaitis, M. E.; Remppe, S. B. Hydration structure and free energy of biomolecularly specific aqueous dications, including Zn^{2+} and first transition row metals. *J. Am. Chem. Soc.* **2004**, *126*, 1285–1289.
- (15) Rogers, D. M.; Beck, T. L. Quasichemical and structural analysis of polarizable anion hydration. *J. Chem. Phys.* **2010**, *132*, 014505.
- (16) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J. P. Anisotropic, polarizable molecular mechanics studies of inter- and intramolecular interactions and ligand-macromolecule complexes. A bottom-up strategy. *J. Chem. Theory Comput.* **2007**, *3*, 1960–1986.
- (17) Cisneros, G. A.; Darden, T. A.; Gresh, N.; Pilmé, J.; Reinhardt, P.; Parisel, O.; Piquemal, J. P. Design Of Next Generation Force Fields From *Ab Initio* Computations: Beyond Point Charges Electrostatics. In *Multi-scale Quantum Models for Biocatalysis* Springer: Netherlands, 2009; Vol. 7, pp 137–172.
- (18) Ren, P. Y.; Ponder, J. W. Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J. Comput. Chem.* **2002**, *23*, 1497–1506.
- (19) de Courcy, B.; Piquemal, J. P.; Gresh, N. Energy Analysis of Zn Polycoordination in a Metalloprotein Environment and of the Role of a Neighboring Aromatic Residue. What Is the Impact of Polarization. *J. Chem. Theory Comput.* **2008**, *4*, 1659–1668.
- (20) Roux, C.; Gresh, N.; Perera, L.; Piquemal, J.; Salmon, L. Binding of 5-phospho-D-arabinonohydroxamate and 5-phos-

- pho-D-arabinonate inhibitors to zinc phosphomannose isomerase from *Candida albicans* studied by polarizable molecular mechanics and quantum mechanics. *J. Comput. Chem.* **2007**, *28*, 938–957.
- (21) Jenkins, L. M. M.; Hara, T.; Durell, S.; Hayashi, R.; Inman, J.; Piquemal, J.; Gresh, N.; Appella, E. Specificity of acyl transfer from 2-mercaptobenzamide thioesters to the HIV-1 nucleocapsid protein. *J. Am. Chem. Soc.* **2007**, *129*, 11067–11078.
- (22) Gresh, N.; Piquemal, J.; Krauss, M. Representation of Zn(II) complexes in polarizable molecular mechanics. Further refinements of the electrostatic and short-range contributions. Comparisons with parallel ab initio computations. *J. Comput. Chem.* **2005**, *26*, 1113–1130.
- (23) Antony, J.; Piquemal, J. P.; Gresh, N. Complexes of thiomandelate and captopril mercaptocarboxylate inhibitors to metallo-beta-lactamase by polarizable molecular mechanics. Validation on model binding sites by quantum chemistry. *J. Comput. Chem.* **2005**, *26*, 1131–1147.
- (24) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. Calculation of protein-ligand binding free energy by using a polarizable potential. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 6290–6295.
- (25) Jiao, D.; Zhang, J. J.; Duke, R. E.; Li, G. H.; Schnieders, M. J.; Ren, P. Y. Trypsin-Ligand Binding Free Energies from Explicit and Implicit Solvent Simulations with Polarizable Potential. *J. Comput. Chem.* **2009**, *30*, 1701–1711.
- (26) Ponder, J. W.; Wu, C. J.; Ren, P. Y.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.
- (27) Grossfield, A. Dependence of ion hydration on the sign of the ion's charge. *J. Chem. Phys.* **2005**, *122*, 024506.
- (28) Jiao, D.; King, C.; Grossfield, A.; Darden, T. A.; Ren, P. Y. Simulation of Ca²⁺ and Mg²⁺ solvation using polarizable atomic multipole potential. *J. Phys. Chem. B* **2006**, *110*, 18553–18559.
- (29) Piquemal, J. P.; Perera, L.; Cisneros, G. A.; Ren, P. Y.; Pedersen, L. G.; Darden, T. A. Towards accurate solvation dynamics of divalent cations in water using the polarizable amoeba force field: From energetics to structure. *J. Chem. Phys.* **2006**, *125*, 054511.
- (30) Bagus, P. S.; Illas, F. Decomposition of the chemisorption bond by constrained variations - Order of the variations and construction of the variational spaces. *J. Chem. Phys.* **1992**, *96*, 8962–8970.
- (31) Stevens, W. J.; Fink, W. H. Frozen fragment reduced variational space analysis of hydrogen-bonding interactions - Application to the water dimer. *Chem. Phys. Lett.* **1987**, *139*, 15–22.
- (32) Jeziorski, B.; Moszynski, R.; Szalewicz, K. Perturbation-theory approach to intermolecular potential-energy surfaces of van der-Waals complexes. *Chem. Rev. (Washington, DC, U. S.)* **1994**, *94*, 1887–1930.
- (33) Becke, A. D.; Edgecombe, K. E. A simple measure of electron localization in atomic and molecular-systems. *J. Chem. Phys.* **1990**, *92*, 5397–5403.
- (34) Silvi, B.; Savin, A. Classification of chemical-bonds based on topological analysis of electron localization functions. *Nature* **1994**, *371*, 683–686.
- (35) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision D.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (36) Ren, P. Y.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.
- (37) Ren, P. Y.; Ponder, J. W. Temperature and pressure dependence of the AMOEBA water model. *J. Phys. Chem. B* **2004**, *108*, 13427–13437.
- (38) Dunning, T. H. Gaussian-Basis Sets for Use in Correlated Molecular Calculations. 1. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (39) Helgaker, T.; Jørgensen, P.; Olsen, J.; Ruud, K.; Andersen, T.; Bak, K. L.; Bakken, V.; Christiansen, O.; Dahle, P.; Dalskov, E. K.; Enevoldsen, T.; Heiberg, H.; Hettema, H.; Jonsson, D.; Kirpekar, S.; Kobayashi, R.; Koch, H.; Mikkelsen, K. V.; Norman, P.; Packer, M. J.; Saue, T.; Taylor, P. R.; Vahtras, O.; Jensen, H. J. A.; Ågren, H. *Dalton, an Ab Initio Electronic Structure Program*, release 1.0, 1997.
- (40) Piquemal, J.; Marquez, A.; Parisel, O.; Giessner-Prettre, C. A CSOV study of the difference between HF and DFT intermolecular interaction energy values: The importance of the charge transfer contribution. *J. Comput. Chem.* **2005**, *26*, 1052–1062.
- (41) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (42) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. *Phys. Rev. B* **1988**, *37*, 785–789.
- (43) Gordon, M. S.; Schmidt, M. W., Advances in electronic structure theory: GAMESS a decade later. In *Theory and Applications of Computational Chemistry, the first forty years*, Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Ed.; Elsevier: Amsterdam, 2005.
- (44) Stevens, W. J.; Basch, H.; Krauss, M. Compact Effective Potentials and Efficient Shared-Exponent Basis-Sets for the 1st-Row and 2nd-Row Atoms. *J. Chem. Phys.* **1984**, *81*, 6026–6033.
- (45) Savin, A.; Nesper, R.; Wengert, S.; Fassler, T. F. ELF: The electron localization function. *Angew. Chem., Int. Ed. Engl.* **1997**, *36*, 1809–1832.
- (46) de Courcy, B.; Pedersen, L. G.; Parisel, O.; Gresh, N.; Silvi, B.; Pilme, J.; Piquemal, J. P. Understanding Selectivity of Hard and Soft Metal Cations within Biological Systems Using

- the Subvalence Concept. 1. Application to Blood Coagulation: Direct Cation-Protein Electronic Effects versus Indirect Interactions through Water Networks. *J. Chem. Theory Comput.* **2010**, *6*, 1048–1063.
- (47) Piquemal, J. P.; Pilme, J.; Parisel, O.; Gerard, H.; Fourre, I.; Berges, J.; Gourlaouen, C.; De La Lande, A.; Van Severen, M. C.; Silvi, B. What can be learnt on biologically relevant systems from the topological analysis of the electron localization function. *Int. J. Quantum Chem.* **2008**, *108*, 1951–1969.
- (48) Pilme, J.; Piquemal, J. P. Advancing beyond charge analysis using the electronic localization function: Chemically intuitive distribution of electrostatic moments. *J. Comput. Chem.* **2008**, *29*, 1440–1449.
- (49) Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. Computational tools for the electron localization function topological analysis. *Comput. Chem.* **1999**, *23*, 597–604.
- (50) Claverie, P. *Intermolecular Interactions: From Diatomics to Biopolymers*; Wiley: New York, 1978; Vol. 1.
- (51) Kutzelnigg, W. The primitive wavefunction in the theory of intermolecular interactions. *J. Chem. Phys.* **1980**, *73*, 343–359.
- (52) Reinhardt, P.; Piquemal, J.-P. *Int. J. Quant. Chem.* **2009**, *109*, 3259–3267.
- (53) Halgren, T. A. Representation of van der Waals (vdW) interactions in molecular mechanics force-fields - potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.* **1992**, *114*, 7827–7843.
- (54) Masia, M.; Probst, M.; Rey, R. On the performance of molecular polarization methods. II. Water and carbon tetrachloride close to a cation. *J. Chem. Phys.* **2005**, *123*, 164505.
- (55) Grossfield, A.; Ren, P. Y.; Ponder, J. W. Ion solvation thermodynamics from simulation with a polarizable force field. *J. Am. Chem. Soc.* **2003**, *125*, 15671–15682.
- (56) Ponder, J. *TINKER: Software Tools for Molecular Design*, version 5.0; Washington University School of Medicine: Saint Louis, MO, 2009.
- (57) Sagui, C.; Pedersen, L. G.; Darden, T. A. Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. *J. Chem. Phys.* **2004**, *120*, 73–87.
- (58) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (59) Bennett, C. H. Efficient Estimation of Free-Energy Differences from Monte-Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245–268.
- (60) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Phys. Rev. Lett.* **2003**, *91*, 140601–1–4.
- (61) Gresh, N. Energetics of Zn²⁺ Binding to a Series of Biologically Relevant Ligands - a Molecular Mechanics Investigation Grounded on Ab-Initio Self-Consistent-Field Supermolecular Computations. *J. Comput. Chem.* **1995**, *16*, 856–882.
- (62) Tiraboschi, G.; Gresh, N.; Giessner-Prettre, C.; Pedersen, L. G.; Deerfield, D. W. Parallel ab initio and molecular mechanics investigation of polycoordinated Zn(II) complexes with model hard and soft ligands: Variations of binding energy and of its components with number and charges of ligands. *J. Comput. Chem.* **2000**, *21*, 1011–1039.
- (63) Stone, A. J. *The Theory of Intermolecular Forces*; Oxford University Press: New York, 1997.
- (64) Zhao, Z.; Rogers, D. M.; Beck, T. L. Polarization and charge transfer in the hydration of chloride ions. *J. Chem. Phys.* **2010**, *132*, 014502.
- (65) Reinhardt, P.; Piquemal, J.; Savin, A. Fragment-Localized Kohn-Sham Orbitals via a Singles Configuration-interaction Procedure and Application to Local Properties and Intermolecular Energy Decomposition Analysis. *J. Chem. Theory Comput.* **2008**, *4*, 2020–2029.
- (66) Marcus, Y. A Simple Empirical-Model Describing the Thermodynamics of Hydration of Ions of Widely Varying Charges, Sizes, and Shapes. *Biophys. Chem.* **1994**, *51*, 111–127.
- (67) Schmid, R.; Miah, A. M.; Sapunov, V. N. A new table of the thermodynamic quantities of ionic hydration: values and some applications (enthalpy-entropy compensation and Born radii). *Phys. Chem. Chem. Phys.* **2000**, *2*, 97–102.
- (68) Mohammed, A. M.; Loeffler, H. H.; Inada, Y.; Tanada, K.-i.; Funahashi, S. Quantum mechanical/molecular mechanical molecular dynamic simulation of zinc(II) ion in water. *J. Mol. Liq.* **2005**, *119*, 55–62.
- (69) D'Angelo, P.; Barone, V.; Chillemi, G.; Sanna, N.; Meyer-Klaucke, W.; Pavel, N. V. Hydrogen and Higher Shell Contributions in Zn²⁺, Ni²⁺, and Co²⁺ Aqueous Solutions: An X-ray Absorption Fine Structure and Molecular Dynamics Study. *J. Am. Chem. Soc.* **2002**, *124*, 1958–1967.
- (70) Obst, S.; Bradaczek, H. Molecular dynamics simulations of zinc ions in water using CHARMM. *J. Mol. Model.* **1997**, *3*, 224–232.
- (71) Kuzmin, A.; Obst, S.; Purans, J. X-ray absorption spectroscopy and molecular dynamics studies of Zn²⁺ hydration in aqueous solutions. *J. Phys.: Condens. Matter* **1997**, *9*, 10065–10078.
- (72) Marini, G. W.; Texler, N. R.; Rode, B. M. Monte Carlo simulations of Zn(II) in water including three-body effects. *J. Phys. Chem.* **1996**, *100*, 6808–6813.
- (73) Yongyai, Y. P.; Kokpol, S.; Rode, B. M. Zinc Ion in Water - Intermolecular Potential with Approximate 3-Body Correction and Monte-Carlo Simulation. *Chem. Phys.* **1991**, *156*, 403–412.
- (74) Born, M. Volumen und Hydratationswärme der Ionen. *Z. Phys. A Hadrons Nuclei* **1920**, *1*, 45–48.
- (75) Garde, S.; Hummer, G.; Paulaitis, M. E. Free energy of hydration of a molecular ionic solute: Tetramethylammonium ion. *J. Chem. Phys.* **1998**, *108*, 1552–1561.
- (76) Rajamani, S.; Ghosh, T.; Garde, S. Size dependent ion hydration, its asymmetry, and convergence to macroscopic behavior. *J. Chem. Phys.* **2004**, *120*, 4457–4466.
- (77) Naor, M. M.; Van Nostrand, K.; Dellago, C. Car-Parrinello molecular dynamics simulation of the calcium ion in liquid water. *Chem. Phys. Lett.* **2003**, *369*, 159–164.
- (78) Badyal, Y. S.; Barnes, A. C.; Cuello, G. J.; Simonson, J. M. Understanding the effects of concentration on the solvation structure of Ca²⁺ in aqueous solutions. II: Insights into longer range order from neutron diffraction isotope substitution. *J. Phys. Chem. A* **2004**, *108*, 11819–11827.
- (79) Jalilvand, F.; Spangberg, D.; Lindqvist-Reis, P.; Hermansson, K.; Persson, I.; Sandstrom, M. Hydration of the calcium ion. An EXAFS, large-angle X-ray scattering, and molecular dynamics simulation study. *J. Am. Chem. Soc.* **2001**, *123*, 431–441.

- (80) Salmon, P. S.; Bellissentfunel, M. C.; Herdman, G. J. The Dynamics of Aqueous Zn-2+ Solutions - a Study Using Incoherent Quasi-Elastic Neutron-Scattering. *J. Phys.: Condens. Matt.* **1990**, *2*, 4297–4309.
- (81) Neely, J.; Connick, R. Rate of water exchange from hydrated magnesium ion. *J. Am. Chem. Soc.* **1970**, *92*, 3476–3478.
- (82) Friedman, H. Hydration complexes - some firm results and some pressing questions. *Chemica Scripta* **1985**, *25*, 42–48.
- (83) Ohtaki, H.; Radnai, T. Structure and dynamics of hydrated ions. *Chem. Rev. (Washington, DC, U. S.)* **1993**, *93*, 1157–1204.
- (84) Helm, L.; Merbach, A. E. Water exchange on metal ions: experiments and simulations. *Coord. Chem. Rev.* **1999**, *187*, 151–181.
- (85) Caminiti, R.; Licheri, G.; Piccaluga, G.; Pinna, G. X-ray-diffraction study of a 3-ion aqueous-solution. *Chem. Phys. Lett.* **1977**, *47*, 275–278.
- (86) Lightstone, F. C.; Schwegler, E.; Allesch, M.; Gygi, F.; Galli, G. A first-principles molecular dynamics study of calcium in water. *ChemPhysChem* **2005**, *6*, 1745–1749.
- (87) Fatmi, M. Q.; Hofer, T. S.; Randolph, B. R.; Rode, B. M. An extended ab initio QM/MM MD approach to structure and dynamics of Zn(II) in aqueous solution. *J. Chem. Phys.* **2005**, *123*, 054514–8.
- (88) Fatmi, M. Q.; Hofer, T. S.; Randolph, B. R.; Rode, B. M. Temperature Effects on the Structural and Dynamical Properties of the Zn(II)-Water Complex in Aqueous Solution: A QM/MM Molecular Dynamics Study. *J. Phys. Chem. B* **2006**, *110*, 616–621.

CT100091J

JCTC

Journal of Chemical Theory and Computation

On the Performances of the M06 Family of Density Functionals for Electronic Excitation Energies

Denis Jacquemin,^{*,†} Eric A. Perpète,[†] Ilaria Ciofini,[‡] Carlo Adamo,^{*,‡}
Rosendo Valero,^{§,⊥} Yan Zhao,^{§,||} and Donald G. Truhlar[§]

Unité de Chimie Physique Théorique et Structurale (UCPTS), Facultés Universitaires Notre-Dame de la Paix, rue de Bruxelles, 61, B-5000 Namur, Belgium, Ecole Nationale Supérieure de Chimie de Paris, Laboratoire Electrochimie et Chimie Analytique, UMR CNRS-ENSCP no. 7575, 11, rue Pierre et Marie Curie, F-75321 Paris Cedex 05, France, Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431, and Commercial Print Engine Lab, HP Laboratories, Hewlett-Packard Co., 1501 Page Mill Road, Palo Alto, California 94304

Received March 2, 2010

Abstract: We assessed the accuracy of the four members of the M06 family of functionals (M06-L, M06, M06-2X, and M06-HF) for the prediction of electronic excitation energies of main-group compounds by time-dependent density functional theory. This is accomplished by comparing the predictions both to high-level theoretical benchmark calculations and some experimental data for gas-phase excitation energies of small molecules and to experimental data for midsize and large chromophores in liquid-phase solutions. The latter comparisons are carried out using implicit solvation models to include the electrostatic effects of solvation. We find that M06-L is one of the most accurate local functionals for evaluating electronic excitation energies, that M06-2X outperforms BHHLYP, and that M06-HF outperforms HF, although in each case, the compared functionals have the same or a similar amount of Hartree–Fock exchange. For the majority of investigated excited states, M06 emerges as the most accurate functional among the four tested, and it provides an accuracy similar to the best of the other global hybrids such as B3LYP, B98, and PBE0. For 190 valence excited states, 20 Rydberg states, and 16 charge transfer states, we try to provide an overall assessment by comparing the quality of the predictions to those of time-dependent Hartree–Fock theory and nine other density functionals. For the valence excited states, M06 yields a mean absolute deviation (MAD) of 0.23 eV, whereas B3LYP, B98, and PBE0 have MADs in the range 0.19–0.22 eV. Of the functionals tested, M05-2X, M06-2X, and BMK are found to perform best for Rydberg states, and M06-HF performs best for charge transfer states, but no single functional performs satisfactorily for all three kinds of excitation. The performance of functionals with no Hartree–Fock exchange is of great practical interest because of their high computational efficiency, and we find that M06-L predicts more accurate excitation energies than other such functionals.

I. Introduction

Time-dependent density functional theory (TD-DFT)^{1–4} is a powerful tool for evaluating properties of electronically

excited states; its predictions are often more accurate than those that can be obtained with other schemes applicable to very large molecules.^{5–24} In addition, medium effects can be readily included in TD-DFT with the help of continuum models^{25–28} (for solvents) or of hybrid quantum mechanical and molecular mechanical (QM/MM) approaches^{29–31} (for biological environments and solid-state catalysts). However, TD-DFT, like DFT for ground electronic states, is in practice applied with approximate density functionals, since an exact functional is unavailable, and many approximate functionals have systematic deficiencies, which have made the predictions less accurate for transitions with Rydberg,^{9,17} long-

* Corresponding author e-mail: denis.jacquemin@fundp.ac.be (D.J.); carlo-adamo@enscp.fr (C.A.).

[†] Facultés Universitaires Notre-Dame de la Paix.

[‡] Ecole Nationale Supérieure de Chimie de Paris.

[§] University of Minnesota.

^{||} Hewlett-Packard Co.

[⊥] Current address: Department of Chemistry, University of Coimbra, 3004-535 Coimbra, Portugal.

range charge-transfer,^{7,17} or double-excitation^{32–34} character in the excited state than for single-excitation valence transitions. It was recently concluded that TD-DFT “still represents the best compromise between accuracy and computational effort. However, large differences in the results are found between the various functionals.”²⁴ Therefore, a well informed choice of the density functional is crucial to generating reliable results. Several extensive tests of various density functionals in the TD-DFT framework have been published; tests have been carried for main-group molecules both in the gas phase and in liquid-phase solutions.^{9,10,12,14–18,20–24} Although the most extensive tests involved more than 20 functionals,²³ the M06 family^{14,35} (M06-L,³⁶ M06,¹⁴ M06-2X,¹⁴ and M06-HF¹⁰) was too new to be included. The present contribution endeavors to fill this lacuna. As in the previous tests,^{9,10,12,14–18,20–24,37} the present results are restricted to the adiabatic linear-response formulation of TD-DFT with density functionals independent of frequency and current, and in particular the adiabatic approximation implies that the functionals developed for ground-state applications are used without change.

In addition to specific applications, the performance of the functionals of the M06 family has been systematically appraised for numerous properties including thermochemistry,^{14,38–42} reaction barriers,^{14,39,41,43,44} catalysis,^{45–48} structural features,^{14,49–52} spin-state energetics,^{49,53} vibrational frequencies and intensities,^{14,54,55} noncovalent interactions,^{14,39,50,51,56–61} and NMR shieldings and related properties.^{62–65} In most cases, the functionals of the M06 family have been found to be relatively broadly accurate and among the most accurate of their respective categories; in particular, M06-L is a very effective local functional (by which we mean a functional that depends on local values of the densities and occupied spin-orbitals (of the noninteracting reference state) and their local derivatives but does not involve an integral over all space as in the Hartree-Fock exchange operator), and the other three are very effective hybrid meta functionals (where “hybrid” denotes the inclusion of Hartree-Fock exchange, and “meta” denotes the inclusion of kinetic energy density, which depends on local derivatives of the spin-orbitals). The investigations in the TD-DFT framework are sparser, but encouraging. One set of tests¹⁰ of M06-HF and six other functionals for main-group excitation energies involved 20 valence excitations, 20 Rydberg-state excitations, and three charge transfer excitations. A later test extended this to M06-L, M06, and M06-2X and 12 older functionals; this test involved 25 valence excitations, 20 Ry excitations, and three charge transfer excitations.¹⁴ In the former study,¹⁰ M06-HF was third best for Rydberg states and best for charge transfer states, but performed poorly for valence excitations. Weighting the three classes of functionals equally, though, it was the best of the seven functionals tested. For these same excitations, weighting the three classes of excitations equally, the subsequent study¹⁴ found M06-HF was best followed by M05-2X (a precursor of M06-2X) and M06-2X. Omitting charge transfer excitations, these three functionals were respectively fifth, second, and third best, out of 16. The 16 density functionals in this study were also applied¹⁴ to five

excitation energies of neutral and cationic metal atoms (including two main-group cases); M06-L and M06 had the third and fourth lowest mean unsigned error for these. In a third systematic study,³⁹ M05-2X, the four members of the M06 family, M08-HX and M08-SO (which are later versions of M06-2X), and six older functionals were applied to nine multiplicity-changing excitation energies; M08-HX, M08-SO, and M06-2X had respectively the first, fourth, and fifth lowest mean unsigned errors, out of 13 functionals tested. One would not necessarily always want to use the functional that predicts, on average, the most accurate excitation energies; in many cases where excitation energies are important, one also needs to accurately model noncovalent interactions and/or barrier heights on the ground potential energy surface, so a broadly accurate functional with good performance for spectroscopy (even if not the best for excitation energies) may be preferable.

In order to more completely evaluate the behavior of the M06 family for the prediction of vertical excitation energies, in this paper, we will test its performance using a variety of databases, designed to include various types of transitions, ranging from valence excitations to charge transfer (CT) and Rydberg states. More specifically, we will consider five databases: two large databases taken from the previous most extensive study of functional performances²³ and three smaller databases covering also CT and Rydberg transitions.¹⁴ The two large databases are called VT and VE to denote “versus theory” and “versus experiment,” respectively.

For the VT tests, we compare TD-DFT results to accurate wave function values for the same transition; in particular, we use data proposed in the recent publications of Thiel and co-workers,^{18,32} in which multistate complete-active-space second-order perturbation theory (MS-CASPT2) and coupled cluster (CC2 and CC3) vertical transition energies were reported for 28 small molecules. This data set was also employed by Goerigk et al. in a study²¹ of doubly hybrid functionals. These VT comparisons entail little ambiguity but have the consequence that only a small and restricted group of molecules (those for which reliable benchmark results are affordable) can be examined. Therefore, in the VE tests, typical families of organic dyes (see Figure 1) encompassing different types of transitions ($n \rightarrow \pi^*$, $\pi \rightarrow \pi^*$, and $\sigma \rightarrow \pi^*$; delocalized and localized) in neutral and charged molecules have been included. Although tests against diverse experimental data are of the greatest importance for validation of theoretical approximations, potential drawbacks of such studies include the difficulty of emulating the environment of the chromophoric molecule under the experimental conditions and sometimes of assigning the transition corresponding to the reported data. One can turn the former issue, namely, environmental effects, into an advantage by using the comparisons as a combined test of density functional approximations and solvation treatments, but it does make a conclusion about the quality of individual density functionals less reliable since it is possible that the best performance could be achieved by a cancellation of errors between the description of the excited state and the treatment of environmental effects. We will minimize the latter issue by choosing molecules where we believe the assignment of the transition or transitions in

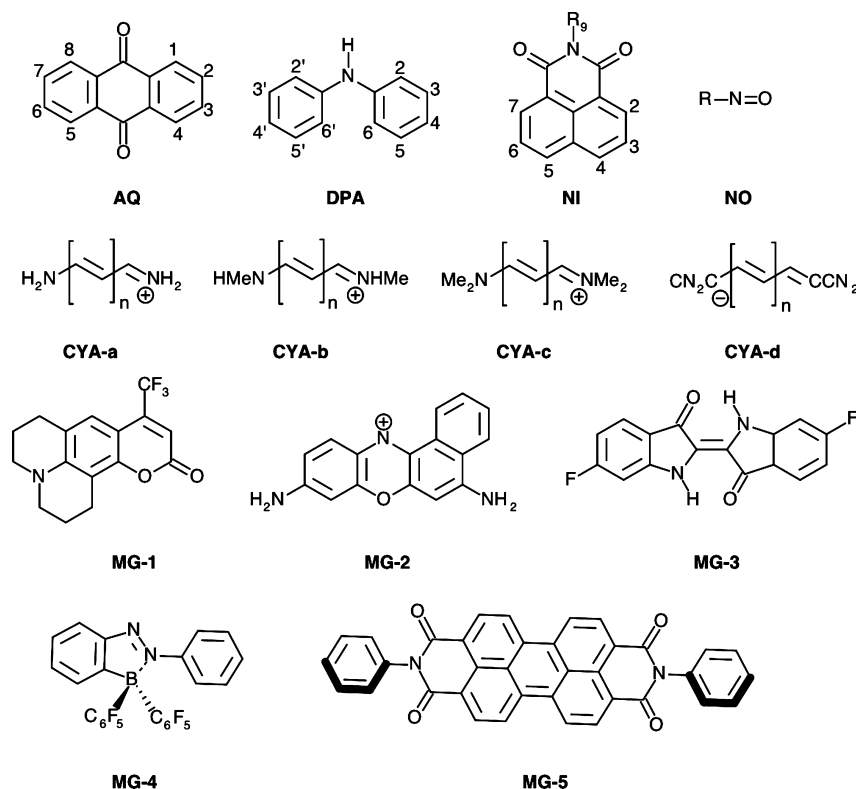


Figure 1. Molecules in the VE set.

question can be made with reasonable confidence. To organize the presentation, the VE database is broken into six subsets.

The transitions in the VT set are all single-excitation valence excitations, and the transitions in the VE set are also predominantly single-excitation valence excitations, although (as discussed below) some of them possess a partial charge transfer character. Other studies, though, showed that one does not draw the same conclusions for Rydberg states,^{10,24} which comprise a large part of the higher-energy spectrum, and for charge transfer states.¹⁰ Two of the smaller databases used here complement the VT and VE databases in that one contains 20 Rydberg-state transitions and the other contains three dominantly charge transfer excitations. These databases are denoted RES20 and CTES3, respectively.¹⁴ RES20 contains experimental data for 20 Rydberg-state transitions of N₂, CO, and HCHO. CTES3 contains theoretical data for NH₃•••F₂ at 6 Å and C₂H₄•••C₂F₄ at 8 Å and experimental data for tetracene. The other small database, VES20, contains 20 energies of valence excited states of N₂, CO, HCHO, and tetracene.^{10,14}

II. Methodology

All calculations have been performed with the *Gaussian* suite of programs, using both standard versions and development versions.^{66–69} All four functionals of the M06 family have been used: M06-HF,¹⁰ M06-L,³⁶ M06,¹⁴ and M06-2X.¹⁴ M06-L is a local meta-GGA functional. Note that, for an open-shell system, “local” denotes that it depends on the local up-spin and down-spin densities and the magnitudes of their gradients and on the local up-spin and down-spin kinetic energy densities (which depend on the self-consistent-field occupied spin-orbitals, which are themselves formally

functionals of the densities). The other three functionals are global-hybrid meta-GGA functionals, where “global-hybrid” denotes the inclusion of a certain percentage (*X*) of Hartree–Fock exchange, and “hybrid meta” denotes that the functionals depend on all the variables of a meta-GGA as well as containing Hartree–Fock exchange (which is computed from the self-consistent-field occupied spin-orbitals). The percentages of Hartree–Fock exchange are 27% for M06, 54% for M06-2X, and 100% for M06-HF. Further details and discussion of the functionals are given elsewhere.^{10,14,35–37}

II.A. Small Molecules: the VT Set. In building the present VT training set, we start with the work of Thiel and co-workers.^{18,32} In their first paper³² (which we will label ES1), they made best estimates for 104 singlets and 63 triplets, out of a total of 223 states (152 singlets and 71 triplets) considered. The best estimates are sometimes from their CC3 calculations, sometimes from their MS-CASPT2 calculations with empirical shifts based on ionization potentials and electron affinities and sometimes from the literature. In a following article¹⁸ (which we label ES2), they provided MS-CASPT2 results for 146 singlets and best estimates for 103 singlets (omitting the ¹B_{3g} state of s-tetrazine from ES1)—in 20 of these cases, they changed the ES1 best estimates, although only by small amounts. The changes are due to relativistic effects, and the mean absolute difference of the MS-CASPT2 estimates of ES1 and ES2 is only 0.01 eV. Here, as in a previous paper,²³ we use the 103 singlets for which ES2 presents best estimates (column 6 of their Table 1); these are all valence transitions (i.e., excited states that are primarily of Rydberg or charge transfer character are not included).

In order to allow consistent comparisons with the results of Thiel's group, we have employed the same basis set (TZVP) and ground-state geometry (MP2/6-31G(d), given elsewhere³² in Cartesian coordinates) that they used^{18,32} for MS-CASPT2 calculations. Note, however, that the "best theoretical estimates" do not always correspond to this geometry or basis set. To provide a comparison with a more consistent choice of geometry and basis set, we also compare to the previously reported³² MS-CASPT2/TZVP results. Although these theoretical values are not completely converged, they are plausibly close enough to the theoretical limit for vertical gas-phase transitions that they may serve as benchmarks for the present work, whose goal is to test the exchange-correlation functionals. This second comparison is not completely independent of the first since some of the best estimates are actually MS-CASPT2/TZVP results.

II.B. The VE Set. For the VE set, we followed the same procedures as in recent tests of other density functionals for dye molecules.^{16,19,70} In this approach, the ground-state structures are first optimized at the PBE0^{71,72} level using the 6-311G(d,p) basis set and the PCM ground-state solvation model²⁷ to simulate bulk liquid solvent effects. Subsequent vibrational analysis—step two—allows confirmation of whether the computed structure is a local minimum of the free energy surface. PBE0 is reasonably accurate for structural parameters of organic molecules,⁷³ and the use of the same PBE0 geometries for all of the liquid-phase tests has the effect that the comparisons of density functionals are not complicated by differences in the geometrical parameters;^{16,74} it also provides consistency with previous work.²³ In the third and final step, the vertical transition energies to the first few valence excited states are calculated using TD-DFT with each of the four density functionals and using the PCM model in a nonequilibrium absorption formulation^{26,27} for inclusion of electrostatic solvent effects. We use the 6-311+G(2d,p) basis set for the TD-DFT calculations on the VE molecules; a summary of tests showing that this basis set is adequate for the kinds of low-lying excited states under consideration here has been provided previously.²³ It is worth noting that the PCM parameters used in the calculations (such as the use of UAKS or UA0 radii or presence or absence of smoothing spheres in defining the solute cavities) vary from one family of dye to another and also depend on the functional (in part because the M06 calculations were carried out with a later version of the code), but the liquid-phase geometries are close enough to the gas-phase ones in most cases that this variation should not be significant enough to affect our conclusions.

The PCM model for electronic spectroscopy includes electrostatic effects of the medium, including the electronic polarizability of the solvent for absorption spectra, but it neglects the difference in dispersion interactions of the solvent with the ground and excited states, and it does not include hydrogen bonding effects beyond their bulk-electrostatic component and so is less accurate for protic solvents. The PCM model employed here could fail when specific solvent–solute interactions take place or when the molecular dipole moment is very different in the ground to the excited states.⁷⁵ The solvents for the VE data used in

this article are benzene (Benz), cyclohexane (CH), chloroform (CHL), dichloroethane (DCE), dichloromethane (DCM), diethyl ether (DEE), dioxane (Diox), ethanol (EtOH), heptane (Hept), hexane (Hex), methanol (MeOH), 2-methylbutane (2MPB), toluene (Tol), and water (Wat). One example of an estimate of the size of the neglected effects on excitation energies is a study of the $n \rightarrow \pi^*$ excitation of acetone in nine solvents, where dispersion effects were estimated to range from 0.07 to 0.09 eV and specific hydrogen bonding effects were estimated to range from 0 to 0.16 eV.⁷⁶ When the errors in the predicted excitation energies of the approximate density functionals are larger than these omitted effects and larger than the errors due to the uncertainties in the included bulk electrostatic effects (we do not have a quantitative estimate of the size of the uncertainties in electrostatics, but they are probably also on the order of 0.1–0.15 eV), we can draw useful conclusions about the density functionals from the comparisons to liquid-phase experimental data.

In principle, one should compare theoretical 0–0 transitions to experimental 0–0 transition energies, but since the latter are usually not available, we compare theoretical vertical transition energies to transition energies calculated from experimental^{77–106} λ_{\max} values, which entails an unknown but probably not insignificant error.^{23,107,108}

Throughout the discussion of the VT molecules, mean absolute deviations (MADs) from the experiment are calculated for the four functionals of the M06 family and compared to those for functionals not in the M06 family and sometimes also to wave function results obtained by time-dependent Hartree–Fock^{5,109} (HF) theory. The latter are calculated from results presented previously^{23,70} for the subsets of cases under discussion in each case.

The last part of the VE data set is a set of five large chromophores (MG-1 to MG-5, Figure 1) for which Goerigk et al.²¹ estimated gas-phase vertical excitation energies from experimental liquid-phase 0–0 transition energies. We compare to the liquid-phase 0–0 data for these five molecules—not to the gas-phase estimates—because medium effects are included in our approach. In particular, we applied our VE methodology to these molecules; i.e., the structures provided by Goerigk et al.²¹ have been reoptimized at the PCM-PBE0/6-311G(d,p) level and vertical (nonequilibrium) PCM-TD-DFT/6-311+G(2d,p) excitation energies have been calculated in the liquid.

In a break with the above, for five of the nitroso dyes in the VE molecule set, we will compare to gas-phase rather than liquid-phase spectra.

II.C. Rydberg and Charge Transfer Excitations. The three small databases, RES20 (Rydberg states), CTES3 (long-range charge transfer excitations), and VES20, are taken from previous work without change.^{10,14}

III. VT Benchmarks

The transition energies obtained for the VT set are listed in Table 1. Discussions of the accuracy obtained for each state by standard GGA (BP86^{110,111}),¹⁸ global hybrids (B3LYP^{112,113} and BHHLYP¹¹⁴),¹⁸ doubly hybrid functionals (B2-LYP, B2GP-LYP, B2-PLYP, and B2GP-PLYP),²¹ and range-

Table 1. VT Test Set: Gas-Phase Electronic Excitation Energies (eV) of Singlet States of Small Molecules^a

molecule	state	M06-L	M06	M06-2X	M06-HF	BE ^b	MS-CASPT2 ^c
ethene	B _{1u} (π)	7.92	7.49	7.80	7.69	7.80	8.54
butadiene	B _u (π)	5.78	5.64	5.97	6.09	6.18	6.47
	A _g (π)	6.67	6.88	7.54	7.88	6.55	6.62
hexatriene	A _g (π)	5.34	5.80	6.58	7.19	5.09	5.42
	B _u (π)	4.67	4.63	4.95	5.15	5.10	5.31
octatetraene	A _g (π)	4.42	4.96	5.76	6.41	4.47	4.64
	B _u (π)	3.97	3.99	4.29	4.53	4.66	4.70
cyclopropene	B ₁ (σ)	6.70	6.33	6.39	6.17	6.76	6.76
	B ₂ (π)	6.43	6.16	6.55	6.64	7.06	7.06
cyclopentadiene	B ₂ (π)	5.05	4.88	5.25	5.36	5.55	5.51
	A ₁ (σ)	6.40	6.53	7.07	7.69	6.31	6.31
norbonadiene	A ₂ (π)	4.82	4.78	5.15	5.24	5.34	5.34
	B ₂ (π)	5.36	5.55	6.04	6.30	6.11	6.11
benzene	B _{2u} (π)	5.41	5.30	5.57	5.77	5.08	5.04
	B _{1u} (π)	6.10	5.87	6.40	6.62	6.54	6.42
	E _{1u} (π)	7.20	6.94	7.20	7.21	7.13	7.13
	E _{2g} (π)	8.65	8.88	9.65	10.23	8.41	8.18
naphthalene	B _{3u} (π)	4.38	4.38	4.64	4.86	4.24	4.24
	B _{2u} (π)	4.24	4.28	4.73	5.08	4.77	4.77
	A _g (π)	6.11	6.13	6.55	6.94	5.87	5.87
	B _{1g} (π)	5.36	5.67	6.27	6.54	5.99	5.99
	B _{3u} (π)	5.96	5.86	6.11	6.21	6.06	6.06
	B _{2u} (π)	6.06	6.00	6.45	6.68	6.33	6.33
	B _{1g} (π)	6.36	6.17	6.66	7.42	6.47	6.47
	A _g (π)	6.87	6.89	7.67	8.08	6.67	6.67
furan	B ₂ (π)	6.29	6.03	6.37	6.47	6.32	6.39
	A ₁ (π)	6.66	6.68	7.14	7.54	6.57	6.50
	A ₁ (π)	8.43	8.09	8.40	8.40	8.13	8.17
pyrrole	A ₁ (π)	6.50	6.48	6.90	7.27	6.37	6.31
	B ₂ (π)	6.48	6.24	6.62	6.77	6.57	6.33
	A ₁ (π)	8.11	7.80	8.11	8.14	7.91	8.17
imidazole	A ⁰ (π)	6.50	6.34	6.75	6.97	6.19	6.81
	A''(n)	6.37	6.36	6.77	6.57	6.81	6.19
	A ⁰ (π)	7.05	6.94	7.39	7.72	6.93	6.93
pyridine	B ₁ (n)	4.76	4.72	4.88	4.68	4.59	5.17
	B ₂ (π)	5.51	5.40	5.66	5.84	4.85	5.02
	A ₂ (n)	4.92	5.05	5.53	6.00	5.11	5.51
	A ₁ (π)	6.30	6.09	6.61	6.83	6.26	6.39
	A ₁ (π)	7.42	7.21	7.50	7.54	7.18	7.46
	B ₂ (π)	7.39	7.18	7.48	7.57	7.27	7.27
pyrazine	B _{3u} (n)	3.90	3.87	3.99	3.80	3.95	4.12
	B _{2u} (π)	5.40	5.26	5.52	5.67	4.64	4.85
	A _u (n)	4.47	4.61	5.04	5.51	4.81	4.70
	B _{2g} (n)	5.55	5.48	5.66	5.27	5.56	5.68
	B _{1u} (π)	6.51	6.28	6.78	6.97	6.58	6.89
	B _{1g} (n)	6.13	6.39	7.15	8.13	6.60	6.41
	B _{2u} (π)	7.83	7.69	8.04	8.20	7.60	7.66
	B _{1u} (π)	7.77	7.57	7.90	7.92	7.72	7.79
pyrimidine	B ₁ (n)	4.15	4.19	4.43	4.44	4.55	4.44
	A ₂ (n)	4.40	4.51	4.92	5.12	4.91	4.80
	B ₂ (π)	5.75	5.65	5.93	6.12	5.44	5.24
	A ₁ (π)	6.58	6.39	6.89	7.10	6.95	6.63
pyridazine	B ₁ (n)	3.54	3.47	3.68	3.50	3.78	3.78
	A ₂ (n)	3.96	4.06	4.56	4.71	4.31	4.31
	A ₁ (π)	5.61	5.52	5.79	5.99	5.18	5.18
	A ₂ (n)	5.34	5.32	5.66	5.99	5.77	5.77
s-triazine	A ₁ '(n)	4.20	4.35	4.87	5.46	4.60	4.60
	A ₂ '(n)	4.43	4.46	4.70	4.73	4.66	4.66
	E''(n)	4.37	4.45	4.79	4.99	4.70	4.70
	A ₂ ⁰ (π)	6.14	6.08	6.37	6.63	5.79	5.79
s-tetrazine	B _{3u} (n)	2.11	2.07	2.28	2.20	2.29	2.29
	A _u (n)	3.20	3.36	3.89	4.27	3.51	3.51
	B _{1g} (n)	4.64	4.64	4.94	4.57	4.73	4.73
	B _{2u} (π)	5.58	5.48	5.75	5.94	4.93	4.93
	B _{2g} (n)	5.13	5.17	5.47	5.32	5.20	5.20
	A _u (n)	4.89	4.88	5.23	5.45	5.50	5.50
formaldehyde	A ₂ (n)	4.23	3.78	3.59	2.99	3.88	3.99
	B ₁ (σ)	9.19	8.67	8.66	8.16	9.10	9.14
	A ₁ (π)	10.61	10.10	9.45	9.33	9.30	9.32
acetone	A ₂ (n)	4.63	4.30	4.10	3.35	4.40	4.44
	B ₁ (σ)	8.61	8.50	8.58	8.11	9.10	9.14
	A ₁ (π)	9.05	9.00	8.91	8.96	9.40	9.32

Table 1. Continued

molecule	state	M06-L	M06	M06-2X	M06-HF	BE ^b	MS-CASPT2 ^c
<i>p</i> -benzoquinone	B _{1g} (<i>n</i>)	2.22	2.48	2.67	2.38	2.76	2.76
	A _u (<i>n</i>)	2.37	2.65	2.85	2.54	2.77	2.77
	B _{3g} (π)	3.61	3.78	4.25	4.74	4.26	4.26
	B _{1u} (π)	4.69	4.89	5.24	5.60	5.28	5.28
	B _{3u} (<i>n</i>)	4.89	5.52	6.36	6.88	5.64	5.64
	B _{3g} (π)	6.46	6.65	7.23	7.86	6.96	6.96
formamide	A''(<i>n</i>)	5.87	5.48	5.37	4.85	5.63	5.63
	A ⁰ (π)	8.02	7.90	8.69	7.68	7.39	7.39
acetamide	A''(<i>n</i>)	5.84	5.54	5.43	4.85	5.69	5.69
	A ⁰ (π)	7.60	7.54	7.97	7.66	7.27	7.27
propamide	A''(<i>n</i>)	5.87	5.57	5.47	4.89	5.72	5.72
	A ⁰ (π)	7.42	7.39	7.62	7.64	7.20	7.20
cytosine	A ⁰ (π)	4.50	4.74	5.03	5.24	4.66	4.67
	A''(<i>n</i>)	4.19	4.80	5.77	5.36	4.87	5.12
	A''(<i>n</i>)	4.88	5.23	5.26	5.49	5.26	5.53
thymine	A ⁰ (π)	5.27	5.55	5.96	6.28	5.62	5.53
	A''(<i>n</i>)	4.48	4.74	4.94	4.61	4.82	4.95
	A ⁰ (π)	4.93	5.05	5.33	5.50	5.20	5.06
	A''(<i>n</i>)	5.24	5.96	6.25	5.85	6.16	6.38
uracil	A ⁰ (π)	5.71	6.19	6.69	6.92	6.27	6.15
	A ⁰ (π)	6.21	6.40	6.78	7.28	6.53	6.53
	A''(<i>n</i>)	4.36	4.67	4.91	4.58	4.80	4.90
	A ⁰ (π)	5.10	5.25	5.51	5.65	5.35	5.23
	A''(<i>n</i>)	5.20	5.87	6.18	5.79	6.10	6.28
	A ⁰ (π)	5.59	6.09	6.56	7.01	6.26	6.15
adenine	A''(<i>n</i>)	5.74	6.30	6.93	6.94	6.56	6.98
	A ⁰ (π)	6.41	6.61	6.94	7.36	6.70	6.74
	A''(<i>n</i>)	4.64	4.90	5.38	5.67	5.12	5.19
	A ⁰ (π)	5.24	5.27	5.57	5.83	5.25	5.20
	A ⁰ (π)	4.85	5.03	5.43	5.66	5.25	5.29
	A''(<i>n</i>)	5.42	5.54	5.93	6.02	5.75	5.96

^a All density functional results use the same TZVP basis set and the MP2/6-31G(d) geometry as in ref 32. The orbital in parentheses (π , n , or σ) denotes a $\pi \rightarrow \pi^*$, $n \rightarrow \pi^*$, or $\sigma \rightarrow \pi^*$ transition, respectively. ^b The BE values are the “best estimates” from ref 18, that is, either CC3 or MS-CASPT2 calculations with empirical IPEA shifts, or are taken from the literature as specified in Table 1 of ref 18 (column 6). ^c MS-CASPT2/TZVP results from ref 18.

separated hybrids (e.g., LC-BLYP,¹¹⁵ LC- ω PBE,¹¹⁶ and CAM-B3LYP¹¹⁷)²³ have already been given in the literature, and therefore we will not discuss individual transitions in detail here; rather we will discuss typical examples.

For the polyenes (butadiene, hexatriene, and octatetraene), the energy of B_u states is always underestimated by M06-L, M06, and M06-2X, but not by M06-HF, which has the smallest average error. It is encouraging that the M06-HF estimates are more accurate than those of the doubly hybrid functionals²¹ for these cases. The A_g states are poorly described due to their significant double-excitation character. We know that ground-state systems with high multireference character are generally treated better by local exchange than Hartree–Fock exchange, and if one makes an analogy between ground states with significant multireference character and excited states with significant double-excitation character, then it is not surprising that the best results for the A_g states are obtained with M06-L, which has only local exchange. The excitation energies of these states are overestimated by the hybrid functionals with the extent of the overestimation increasing with *X*, leading, for example, to very large errors of ~ 2 eV in the M06-HF calculations on hexatriene and octatetraene.

For benzene and naphthalene, the transition energies obtained with the three hybrids usually follow the trend that a larger *X* yields larger transition energies, although the trend between M06-L and M06 is sometimes an exception. For these two aromatic compounds, the MADs from the bench-

mark values increase with *X*: 0.29 eV with M06-L, 0.33 eV with M06, 0.40 eV with M06-2X, and 0.72 eV with M06-HF.

In the heterocyclic series, the first B_{2u} state of pyrazine and the first $\pi \rightarrow \pi^*$ transition of *s*-triazine are examples of challenging states.²³ For the former, M06 (5.26 eV) is closer to the best estimate (4.64 eV) and MS-CASPT2 results (4.85 eV) than B3LYP (5.37 eV)¹⁸ or PBE0 (5.44 eV),²³ but B2-PLYP is the most accurate of all functionals examined (5.16 eV).²¹ The same ranking in accuracy is obtained for the latter case. For the 20 $n \rightarrow \pi^*$ transitions of the heterocyclic subset, the MAD (using MS-CASPT2/TZVP values as benchmarks) is 0.31 eV for M06-L, 0.26 eV for M06, 0.19 eV for M06-2X, and 0.41 eV for M06-HF. In comparison, PBE0 gives notably smaller deviations for this subset (MAD of 0.13 eV),²³ and the accuracy of doubly hybrid functionals is also superior.²¹

In the series containing ketones, aldehydes, and amines, M06 is the best performing functional of Table 1 (MAD of 0.28 eV); it provides an error similar to that of B3LYP¹⁸ and PBE0.²³

For the four nucleotide bases, which are the largest systems of the VT set, the M06 and M06-2X functionals produce similar average deviations, with MADs of 0.21 and 0.25 eV, respectively, vs the MS-CASPT2 reference. Therefore, the deviations of M06 are similar to those of B3LYP, but M06-2X is almost twice as accurate as BHLYP.

Table 2. Mean Deviations (in eV) of the Density Functional Predictions from the Best Estimates of ref 32^a

functional	X	MSD	MAD	RMSD	R ²	ref
BP86	0	0.44	0.52	0.62	0.92	22
M06-L	0	0.14 (0.18)	0.35 (0.37)	0.42 (0.47)	0.91 (0.93)	this work
B3LYP	20	0.07	0.27	0.33	0.94	22
M06	27	0.12 (0.16)	0.28 (0.31)	0.34 (0.38)	0.95 (0.95)	this work
BHLYP	50	-0.43	0.50	0.62	0.89	22
DFT/MRCI	50	0.13	0.22	0.29	0.96	22
B2-LYP	53	0.45	0.52	0.62	0.90	24 ^b
B2-PLYP	53	0.01	0.18	0.25	0.97	24 ^b
M06-2X	54	-0.23 (-0.18)	0.34 (0.35)	0.46 (0.46)	0.92 (0.92)	this work
M06-HF	100	-0.32 (-0.28)	0.55 (0.56)	0.70 (0.70)	0.83 (0.83)	this work

^a The values in parentheses are deviations from MS-CASPT2/TZVP benchmarks for the same series of states (see Table 1 for both the best estimates and MS-CASPT2/TZVP results). The MSD, MAD, and RMSD are in eV, and R² is the square of the linear correlation coefficient. ^b Note that the values listed for B2-LYP and B2-PLYP have been recalculated from the raw data of ref 21 in order to use the same standard values for all results in the whole table.

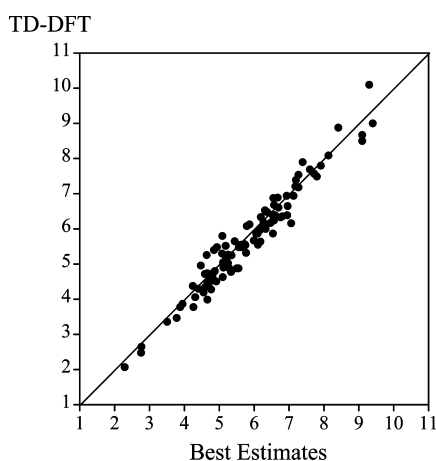


Figure 2. Comparison between TD-M06 predictions and best estimates for vertical transition energies for the full VT set (103 transitions). All values are in eV. The line at 45° corresponds to a perfect match between the two sets of values. For this data set, the MAD of M06 is 0.28 eV.

The mean signed deviations (MSDs), MADs, and root-mean-square deviations (RMSDs) obtained with the M06 family of functionals are compared to previous benchmarks in Table 2, and for the M06 functional, a graphical comparison to the best estimates is given in Figure 2. The squares of the correlation coefficients (R²) obtained by linear fitting, are also reported. Using MS-CASPT2/TZVP instead of “best estimates” as reference values leads to a small increase of the MSD (+0.04 eV) and MAD (+0.02 eV) but does not significantly affect the trends.

Irrespective of the chosen reference, M06 systematically provides the smallest errors and the largest correlation coefficient of the four functionals of the M06 family. The quality of the M06 results for the present tests is similar to that reported previously¹⁸ for B3LYP. Comparing the two functionals with X = 0, we see that M06-L surpasses BP86, with a MSD reduced by a factor of 3 and an RMSD reduced by 0.2 eV! The errors of M06-L do remain sizable, but this meta-GGA reduces the differences with respect to global hybrids, which is interesting from the point of view of the Jacob’s ladder classification¹¹⁸ of functionals since meta-GGAs are on rung 3 and hybrid GGAs and hybrid meta-GGAs are on rung 4.

Moving on to the larger percentages of Hartree–Fock exchange (larger X values), M06-2X has a similar performance to BHLYP, although they have similar percentages; the improvement is pronounced in terms of both correlation and average deviations. M06-HF produces large errors (similar to those of BP86, but with the opposite sign) and a poor correlation: it is unlikely to be of interest for calculations on excited states if only valence excitations are of interest. The DFT/MRCI scheme^{18,119} and the B2-PLYP doubly hybrid functional are significantly more accurate (although also more complicated, especially DFT/MRCI, which attempts to provide a more realistic treatment of doubly excited states) than M06, and B2-PLYP is the most accurate functional tested up to now for this set. This finding is again consistent with the Jacob’s ladder classification, since B2-PLYP is a doubly hybrid functional^{21,120} on rung 5 (the highest rung), whereas all the other functionals with nonzero X in Table 2 are on rung 4. The DFT/MRCI method in Table 2 also uses unoccupied DFT orbitals.

Figure 2 shows a graphical comparison of the M06 excitation energies to the best estimates for the full VT test set.

IV. The VE Database

The molecules belonging to the VE database can be divided into six families: the 9,10-anthraquinones (AQ, Figure 1), the (nitro)-diphenylamines (DPA, Figure 1), the 1,8-naphthalimides (NI, Figure 1), the nitroso dyes (RNO, Figure 1) the cyanines (CYA-x, Figure 1), and the large chromophores (MG-y). This latter is a subset of five dyes (MG1–MG5, Figure 1), recently studied by Goerigk et al.²¹ Table 3 collects the MSDs, MADs, RMSDs, and R² values computed using the four functionals belonging to the M06 family for all these systems, while a detailed list of the computed transition energies for the six families as well as specific discussions can be found in the Supporting Information (Tables SI.1 to Table SI.6 and related text).

IV.A. Anthraquinones (AQ), Diphenylamines (DPA), and Naphthalimides (NI): The ππ*D49 Subset. The AQ, DPA, and NI families are constituted by dyes (AQ and DPA) and fluorophores (NI) largely studied both experimentally^{102,103,121–123} and theoretically,^{22,23,70,124–129,131–134} and taken together, they provided a suitable database of 49 excitation energies (30 AQ, 11 DPA, and 7 NI, respectively),

Table 3. Mean Signed (MSD) and Unsigned (MAD) Deviations (in eV) from Experimental Transitions Computed for the Molecules Belonging to the AQ, DPA, NI, RNO, Cya-x, and MG-y Families Together with the Corresponding RMSD (in eV) and R^2 Values^a

	MSD			
	M06-L	M06	M06-2X	M06-HF
AQ	0.27	-0.03	-0.43	-0.84
DPA	0.45	0.13	-0.33	-0.64
NI	0.19	0.01	-0.30	-0.58
RNO	-0.21	0.20	0.39	1.08
Cya-x	-0.67	-0.55	-0.57	-0.50
MG-y	0.20	0.11	-0.13	-0.31

	MAD			
	M06-L	M06	M06-2X	M06-HF
AQ	0.29	0.11	0.43	0.84
DPA	0.45	0.14	0.33	0.64
NI	0.19	0.08	0.30	0.58
RNO	0.23	0.21	0.39	1.08
Cya-x	0.67	0.55	0.57	0.50
MG-y	0.26	0.18	0.14	0.31

	RMSD			
	M06-L	M06	M06-2X	M06-HF
AQ	0.33	0.13	0.43	0.85
DPA	0.47	0.15	0.35	0.67
NI	0.22	0.10	0.30	0.58
RNO	0.22	0.28	0.41	1.13
Cya-x	0.68	0.56	0.57	0.51
MG-y	0.29	0.19	0.17	0.34

	R^2			
	M06-L	M06	M06-2X	M06-HF
AQ	0.89	0.96	0.98	0.97
DPA	0.91	0.95	0.91	0.78
NI	0.81	0.91	0.96	0.98
RNO	0.96	0.99	0.98	0.81
Cya-x	0.99	0.99	1.00	0.99
MG-y	0.73	0.87	0.94	0.95

^a A detailed list of computed and experimental transition energies is reported in the Supporting Information (Tables SI.1–SI.6).

all of the $\pi-\pi^*$ type, allowing for a robust benchmark to assess density functional performances for this type of transition. The MADs for the Hartree–Fock approximation and 14 density functionals for this data set (hereafter called $\pi\pi^*$ D49) are reported in Table 4. The table also gives references for the density functionals^{10,14,36,71,72,109,110,112,113,135–141} to which we compare.

In all cases, even though the members of the M06 family have different exchange and correlation potentials, as well as different values of X , the computed transition energies of the M06 family perfectly follow the percentage of Hartree–Fock exchange, i.e., $M06-L < M06 < M06-2X < M06-HF$; thus this percentage seems to be the most important parameter, as already observed for $\pi-\pi^*$ excitations.^{70,130} Three other common features for these three families of compounds can also be observed by inspection of Tables 3 and 4, in particular: (i) M06-L underestimates the transition energies but nevertheless outperforms all the previously

Table 4. Mean Unsigned Deviations (eV) from Best Estimates for the $\pi\pi^*$ D49 Data Set

functional	X^a	ref	AQ	DPA	NI	$\pi\pi^*$ D49
M06-HF	100	10	0.84	0.64	0.58	0.75
M05-2X	56	136	0.45	0.40	0.41	0.43
M06-2X	54	14	0.43	0.33	0.30	0.39
HF ^b	100	109	1.15	1.26	0.81	1.13
BMK	42	137	0.31	0.27	0.32	0.30
PBE0	25	71, 72	0.10	0.06	0.11	0.09
B98	21.98	138	0.11	0.12	0.09	0.11
M05	28	135	0.11	0.09	0.12	0.11
B3LYP	20	112, 113	0.13	0.18	0.08	0.14
M06	27	14	0.11	0.14	0.08	0.11
TPSSh	10	139	0.23	0.31	0.08	0.23
M06-L	0	36	0.29	0.45	0.19	0.31
VSXC	0	140	0.35	0.49	0.20	0.36
PBE	0	141	0.47	0.60	0.32	0.48
BLYP	0	110, 112	0.47	0.65	0.34	0.50

^a X denotes percentage of Hartree–Fock exchange. ^b This row (Hartree–Fock) is wave function theory; other rows are density functional theory.

benchmarked GGA functionals (such as PBE and BLYP) and also meta-GGAs (e.g., VSXC); (ii) the performance of M06 makes it one of the best global hybrids for this category of dyes (in the case of the $\pi\pi^*$ D49 data set, a MAD of 0.11 eV is computed for M06 as compared to 0.09 eV for PBE0 and 0.14 eV for B3LYP); (iii) M06-2X and M06-HF predict systematically higher transition energies; (iv) the MAD achieved with M06-HF is much smaller than with TD-HF (0.75 eV versus 1.13 eV for the $\pi\pi^*$ D49). Moreover, the correlation between experimental and theoretical values is usually good ($R^2 = 0.91–0.98$) for the hybrid functionals of the M06 family (with the only noteworthy exception being M06-HF for DPA) but can be significantly lower for M06-L (R^2 is only 0.81 and 0.89, in the case of NI and AQ, respectively). Finally, it is also worthwhile to remember that, in addition to the performance of the M06 functional being very close to the performances of other global hybrids that include a similar amount of Hartree–Fock exchange, fine details of substituent effects are better described by M06. For instance, M06 predicts the correct ordering for the 1,4-OH versus 1-NH₂ substitutions as well as for the 1,2-OH versus 1,8-OH patterns in the AQ family, a feat that neither PBE0 nor range-separated hybrids could achieve.⁷⁰ A more detailed discussion of computed transition energies can be found in the Supporting Information.

IV.B. Nitroso Dyes (NO18) and Cyanines (CYA13) Subsets. Due to the large separation between the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ bands, nitroso derivatives (NO, Figure 1) are well-known $n \rightarrow \pi^*$ chromogens.^{121,122} The UV/vis features of NO dyes have been tackled by some of us in three previous studies of TD-DFT.^{12,23,142} The full list of the transitions (18) computed using the M06 family and a brief comment on their ordering are reported in the Supporting Information (Table SI.4 and related text).

The general trends for the nitroso dyes do not follow the pattern seen in section IV.A. For example, larger percentages of Hartree–Fock exchange generally imply smaller transition energies for nitroso dyes. Consequently, the MSDs all have the opposite sign of those for the AQ dyes (Table 3). M06 again has the smallest MAD and the largest correlation

coefficient of any member of the M06 family. The M06 MAD (0.21 eV) is significantly smaller than for its M05 precursor (0.33 eV), but larger than for PBE0 (0.08 eV). As a group, the M06 family does relatively poorly compared to other density functionals for the nitroso dyes.

Cyanine dyes are charged dyes (both anionic and cationic derivatives are considered) with highly delocalized structures. Although the four series treated in the present contribution (CYA-x, Figure 1) belong to the streptocyanine subcategory, other structures (like malachite green or nile blue) have similar electronic characteristics.⁹⁹ Due to the strong multideterminantal nature of the states of these dyes,¹⁴³ TD-DFT does not correctly predict the absolute variations of the transition energies as chain length increases. This is true with conventional hybrids,^{144,145} range-separated hybrids,^{23,130} and even doubly hybrid functionals,¹⁴⁶ though the latter provide slightly smaller absolute deviations. Table SI.5, collecting the 13 computed transitions, and Table 3 show that no functional of the M06 family succeeds in improving the usual dreadful errors, and the transition energies are still uniformly overestimated. More positively, one notes excellent correlation coefficients (Table 3) for all four functionals, just as good correlation coefficients can be obtained with other theoretical methods as well.²³ In fact, for CYA-x, the nature of the selected functional appears to be almost irrelevant, although MS-CASPT2 seems capable of mirroring the experimental measurements.¹⁴³ We will return to the classification of these transitions in section V.

IV.C. Large Chromophores (MG-y) Subset. In Table SI.6 (Supporting Information), we report the M06 family results (five transitions) for the set of large dyes recently studied by Goerigk et al.²¹ As explained in the Methodology section, estimates were made for the energies of the experimental vertical transitions, thereby attempting to remove the drawback of comparing calculated vertical transition energies to transition energies corresponding to the wavelength of maximum absorption.

Two challenging cases are discussed in more detail in the Supporting Information, and here we consider average errors for all five large chromophores, Table 3 shows the smallest errors for M06 (MAD of 0.18 eV) and M06-2X (MAD of 0.14 eV); it also shows that M06-L usually underestimates the transition energies, and M06-HF always overestimates them. For the sake of comparison, we have computed a PBE0 MAD (MSD) of 0.14 eV (0.03 eV) for the same set of five large chromophores. This is the same average error as the one obtained for a much larger set of dyes.²³ This implies that the errors obtained for low-lying excited states of organic dyes (the VE set) are smaller (on average) than for high-energy states of small molecules (see the VT set). This MAD (MSD) is substantially smaller than the one reported previously for the same functional: 0.20 eV (0.11 eV),²¹ illustrating that changes in the geometrical parameters (PBE/TZVP²¹ versus PCM-PBE0/6-311G(d,p)) and basis set (TZVP versus 6-311+G(2d,p)) substantially affect the conclusions. The most effective functional was previously²¹ found to be B2GP-PLYP, which is associated with a MAD of 0.16 eV, but this value could also be overestimated due to the testing methodology. Therefore, although we expect a non-negligible

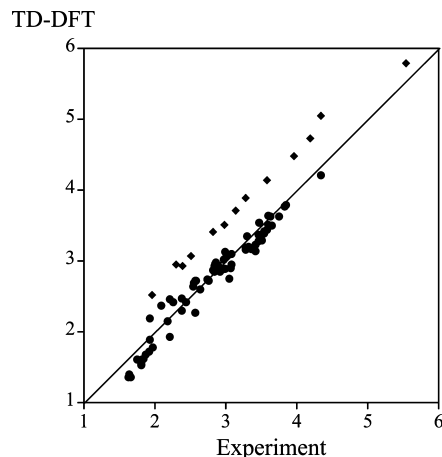


Figure 3. Comparison between experimental and theoretical (M06) transition energies (eV) for the VE set consisting of the combined $\pi\pi^*$ D49, NO18, and CYA13 data sets. The closed diamonds correspond to the CYA-x series. For this data set, the MAD of M06 is 0.20 eV.

improvement by using doubly hybrid functionals, the quantitative extent of this effect remains unsettled for large molecules.

Figure 3 shows a graphical comparison of the M06 predictions to the best estimates for the full VE data set.

V. Including Rydberg and Charge Transfer Excitations in the Assessment

The final classes of data that we consider are for Rydberg and charge transfer excitations. It is important that practical density functionals do not have large errors for these classes of excitations, because in complex molecules many transitions have some Rydberg and/or charge transfer character, and if this kind of excitation is not treated well, some components of the excited state will be misrepresented even when the predominant character of an excitation is valence-like. For example, even to treat the $\pi \rightarrow \pi^*$ excitation of ethylene correctly, it is necessary to treat valence and Rydberg states on an even-handed basis,^{147,148} and the amount of Rydberg character in a given transition can depend strongly on geometry. Charge transfer presents similar difficulties in that the extent of charge transfer covers a very wide range when one surveys a range of molecules.¹⁷ Furthermore, as mentioned at the end of section I, Rydberg states are not included in the test cases considered in both the VT and VE sets, nor is charge transfer character strongly represented in those test cases.

To illustrate the problems encountered in charge transfer states, we first consider a prototype charge transfer case, namely, the C_2H_4 molecule at a fixed distance R from a C_2F_4 molecule. The orientation is shown in Figure 4. We consider two values of R , in particular, 4 Å and 8 Å. At these distances, there is little spatial overlap of the densities of the HOMO and the LUMO, so the lowest excitation may be classified unambiguously as a charge transfer excitation (transitions that may be classified this way due to lack of overlap are called long-range charge transfer). Best estimates are obtained from the wave function calculations of Tawada et al.¹⁴⁹ and Dreuw et

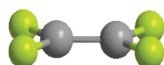
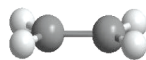


Figure 4. C_2H_4 molecule at a fixed distance R from a C_2F_4 molecule. In the figure, R is 8 Å.

Table 5. Charge Transfer Excitation Energies of $C_2H_4 \cdots C_2F_4$ Separated by 4 and 8 Å and Mean Absolute Deviations from Best Estimates^a

	ref	X	4 Å	8 Å	MAD
TD-PBE0/6-31G(d)	71, 72	25	6.74	7.41	4.60
TD-PBE0/6-31+G(d)	71, 72	25	6.53	7.35	4.74
TD-PBE0/6-31++G(2d,p)	71, 72	25	6.44	7.22	4.85
TD-PBE0 ^b	71, 72	25	6.48	7.26	4.81
TD-M06-HF/6-31G(d)	10	100	10.58	12.62	0.08
TD-M06-HF	10	100	9.30	11.45	1.30
TD-M05-2X	136	56	7.62	9.32	3.21
TD-M06-2X	14	54	7.17	8.89	3.65
TD-HF ^c	109	100	10.36	12.30	0.35
TD-HFLYP	109, 112	100	11.47	12.84	0.48
TD-BMK	137	42	7.41	8.54	3.70
TD-B97-3	151	26.93	6.68	7.45	4.61
TD-B98	138	21.98	6.36	7.04	4.98
TD-M05	135	28	6.50	7.40	4.73
TD-B3LYP	112, 113	20	6.25	6.90	5.10
TD-mPW1PW	152	25	6.50	7.29	4.78
TD-X3LYP	153	21.8	6.30	7.04	5.01
TD-M06	14	27	6.77	7.21	4.69
TD-M06-L	36	0	5.43	5.70	6.11
TD-BLYP	110, 112	0	5.06	5.26	6.52
TD-CAM-B3LYP	117	19–65 ^d	7.11	9.01	3.62
TD-LC- ω PBE(20)	23	0–100 ^d	6.48	8.42	4.23
TD-LC- ω PBE	116	0–100 ^d	9.08	10.81	1.73
MS-CASPT2(4e/4o) ^{c,e}	154	100	9.15	10.09	2.06
best estimate ^f			10.72	12.63	0

^a Complex constructed taking the experimental geometries of the monomers. ^b In this table, if the basis set is not indicated, it is 6-311+G(2d,p). ^c Wave function theory (other entries are density functional theory). ^d The lower end of the range applies at zero interelectronic distance, and the upper end of the range applies at infinite interelectronic distance. ^e 4e/4o denotes four electrons in four active orbitals. These calculations were performed with MOLCAS and are based on four-state-averaged CASSCF orbitals. ^f For $R = 8.00$ Å, the best estimate comes from the SAC–CI results of Tawada et al.¹⁴⁹ The SAC–CI results are only available for $R \geq 5$ Å, but for 5 and 6 Å, the difference between the SAC–CI excitation energies and the CIS results in Figure 3 of Dreuw et al.¹⁵⁰ is constant at 0.53 eV. With this difference, the SAC–CI value at $R = 5$ Å of 11.49 eV and the CIS difference between $R = 4$ and 5 Å of 0.77 eV, we obtain a best estimate of 10.72 eV at 4 Å.

al.,¹⁵⁰ as explained in a footnote to Table 5, which also shows results for PBE0 with four basis sets, M06-HF with two, and—with one basis set—several other density func-

Table 6. Mean Absolute Deviations from Best Estimates for RES20 Database of Rydberg Excitation Energies^a

	ref	X	MAD ^a
TD-M06-HF	10	100	0.43
TD-M05-2X	136	56	0.31
TD-M06-2X	14	54	0.35
TD-HF	109	100	1.18
TD-HFLYP	109, 112	100	1.72
TD-BMK	137	42	0.35
TD-BHLLYP	114	50	0.17
TD-B97-3	151	26.93	0.78
TD-PBE0	71, 72	25	0.86
TD-B98	138	21.98	0.88
TD-M05	135	28	1.10
TD-B3LYP	112, 113	20	0.67
TD-mPW1PW	152	25	0.84
TD-X3LYP	153	21.8	0.99
TD-O3LYP	112, 155	11.61	1.55
TD-M06	14	27	1.67
TD-M06-L	36	0	1.62
TD- τ -HCTHhyb	156	15	1.08
TD- τ -HCTH	156	0	1.69
TD-TPSS	157	0	1.72
TD-BP86	110, 111	0	1.85
TD-BLYP	110, 112	0	2.00
TD-OLYP	112, 155	0	2.13
TD-SVWN5	158	0	1.77
TD-CAM-B3LYP	117	19–65	0.50
TD-LC- ω PBE(20)	23	0–100	1.14
TD-LC- ω PBE	116	0–100	0.15
TD-LC-BLYP	159	0–100	0.21
TD-LC-OLYP	23, 159	0–100	0.22
TD-LC-PBE	23, 159	0–100	0.34
TD-LC- τ -HCTH	23, 159	0–100	0.92
TD-LC-TPSS	23, 159	0–100	0.48

^a Augmented Sadlej pVTZ basis set.

tional approximations^{14,23,36,109,110,112,113,116,17,135–138,151–154} and two wave function calculations. The basis set dependence is small compared to the errors for PBE0. The MADs from the best estimates are also shown. These results are typical for long-range charge transfer excitations; only functionals with both electron correlation and 100% Hartree–Fock exchange at all values of the electronic coordinates show useful accuracy. Even the long-range corrected TD-LC- ω PBE and TD-LC- ω PBE(20) methods, which have 100% Hartree–Fock exchange in the limit of large interelectronic separation, have large errors. These results are consistent with earlier studies showing how difficult it is to get useful results for long-range charge transfer.^{10,14,17,145}

We next illustrate a similar—but not quite as dramatic—problem with Rydberg transitions. For these calculations, we used the RES20 database of 20 Rydberg-state excitation energies^{10,14} and the augmented Sadlej pVTZ basis set, and the mean unsigned errors are in Table 3. Again, we compare the results obtained with the M06 family to calculations with several other methods.^{23,71,72,109–114,116,117,135–138,151–153,155–159}

Table 6 shows that all functionals with less than 42% Hartree–Fock exchange give very poor results for Rydberg states. M06-HF, M05-2X, M06-2X, BMK, LC- ω PBE, LC-BLYP, LC-OLYP, and LC-PBE give MADs from the best results of less than 0.45 eV, but only LC- ω PBE, LC-BLYP, and LC-OLYP give MADs of 0.22 eV or less.

On the basis of the above considerations, in order to consider valence, Rydberg, and long-range charge transfer

Table 7. Mean Absolute Deviations (eV) from Best Estimates for Combined Data Sets

functional	X^a	VT103	$\pi\pi^*D49$	NO18	VES20	VES190	RES20	VRES210	CYA13	CTES3	ES226	BES226
M06-HF	100	0.55	0.75	1.08	0.71	0.67	0.39	0.64	0.50	0.09	0.63	0.54
M05-2X	56	0.39	0.43	0.51	0.37	0.41	0.31	0.40	0.66	2.42	0.44	0.53
M06-2X	54	0.34	0.39	0.39	0.34	0.36	0.35	0.36	0.57	2.46	0.40	0.50
HF ^b	100	1.05	1.13	0.15	1.08	0.99	1.18	1.01	1.08	0.99	1.01	1.05
BMK	42	0.34	0.30	0.19	0.30	0.31	0.35	0.32	0.68	3.10	0.37	0.53
PBE0	25	0.24	0.09	0.08	0.29	0.19	0.86	0.26	0.63	4.08	0.33	0.63
B98	21.98	0.25	0.11	0.07	0.24	0.20	0.92	0.26	0.61	4.25	0.34	0.65
M05	28	0.30	0.11	0.33	0.29	0.25	1.16	0.34	0.61	4.12	0.40	0.73
B3LYP	20	0.27	0.14	0.06	0.28	0.22	1.07	0.30	0.59	4.44	0.37	0.70
M06	27	0.28	0.11	0.21	0.24	0.23	1.67	0.36	0.55	4.11	0.42	0.83
TPSSh	10	0.30	0.23	0.12	0.24	0.26	1.33	0.36	0.62	4.93	0.44	0.82
M06-L	0	0.35	0.31	0.23	0.32	0.33	1.62	0.45	0.67	5.44	0.53	0.96
VSXC	0	0.39	0.36	0.17	0.27	0.35	1.64	0.47	0.65	5.63	0.55	0.98
PBE	0	0.53	0.48	0.15	0.32	0.46	1.95	0.60	0.51	5.86	0.67	1.10
BLYP	0	0.54	0.50	0.14	0.35	0.47	2.00	0.62	0.50	5.85	0.68	1.11

^a X denotes percentage of Hartree–Fock exchange. ^b This row (Hartree–Fock) is wave function theory; other rows are density functional theory.

in our assessment, results from different data sets are combined in Table 7. To this end, three small data sets (VES20, RES20, and CTES3—with all results taken from a previous paper,¹⁴ which can be consulted for details such as basis sets, geometries, and sources of accurate data) are compared to several data sets from this paper, namely, VT103, which consists of the results for the 103 excitations energies in the VT set; $\pi\pi^*D49$, which is defined above (recall that it consists of 49 $\pi \rightarrow \pi^*$ excitations of various neutral dyes); NO18, which consists of 18 $n \rightarrow \pi^*$ transitions of neutral nitroso dyes; and CYA13, which consists of 13 transitions involving highly multiconfigurational states of charged cyanines.

Table 7 does not include the five large chromophores (MG-y family, Table SI.6, Supporting Information) because we do not have results for those molecules for most of the density functionals included in Table 7. The VES190, VRES210, ES226, and BES226 columns of Table 7 are explained below.

First, we notice the difference between the trends in the CTES3 column and the VT103, $\pi\pi^*D49$, NO18, and VES20 columns of Table 7. For the long-range charge transfer states of CTES3, the errors are smallest for M06-HF, with $X = 100$, whereas for the VT103, $\pi\pi^*D49$, NO18, and VES20 databases of valence excitations, the errors are smallest for $X = 10$ –28. The CYA13 column shows a trend more in line with CTES3 than with the valence-excitation databases, although the trend is not as pronounced as for the long-range charge transfer excitations of CTES3. On the basis of this observation, we will classify the cyanine data as charge transfer excitations for the rest of this discussion.

We next defined VES190 as a database of the 190 valence excitations in VT103, $\pi\pi^*D49$, NO18, and VES20, and Table 7 shows the mean errors over all 190 data. The five best performing functionals are the ones with $X = 20$ –28, and they all have MADs in the range 0.19–0.25 eV. TPSSh, BMK, M06-L, and M06-2X are the closest trailers, with MADs in the range 0.26–0.36 eV. Adding in the Rydberg transitions of RES20 makes database VRES210 with 210 valence and Rydberg excited states; for these data, PBE0 and B98 have the best performance with MAD = 0.26 eV. M06 and M06-2X both have a MAD of 0.36 eV. Finally,

we add in the 16 charge transfer excitations of CYA13 and CTES3, and we obtain the largest data set, ES226, with 226 excited states. PBE0 and B98 remain the best performers, with MAD = 0.33–0.34 eV; the best performers in the M06 family are M06-2X and M06, with MADs of 0.40 and 0.42 eV. The 190:20:13 relative weighting of valence/Rydberg/charge-transfer excitations in ES226 (equivalent to 84:9:7) is very arbitrary. An alternative method of gaining an overall perspective would be, for example, to use a 50:25:25 weighting. We do this by first combining CYA13 and CTES3 into CTES16 and then compute a MAD for “balanced ES226” as follows:

$$\text{MAD}(\text{BES226}) = 0.50 \times \text{MAD}(\text{VES190}) + 0.25 \times \text{MAD}(\text{RES20}) + 0.25 \times \text{MAD}(\text{CTES16})$$

Such an assessment is shown in the last column of Table 7. Although it is equally as arbitrary as the raw average over the 226 molecules in ES226, it might be a more useful test when one considers a set of states having all three kinds of character. The table shows that the high- X functionals are now the best, followed by the mid- X functionals. Therefore, each potential user of the methodology must first determine whether Rydberg and/or charge transfer excitations are an important component of the transition set being studied, and the decisions on the usefulness of TD-DFT and the optimal functional depend strongly on that consideration. The bottom line is that the current situation is an unsatisfactory state of affairs because no single local or global hybrid functional is reasonably accurate for all three classes of excitation.

This is illustrated in Figure 5 where the MADs computed using the functionals belonging to the M06 family and four different combined data sets (VES190, VRES210, ES226, and BES226) are compared to the performances of other local and global hybrids functionals.

VI. Conclusions

It is important to validate practical density functional approximations in order to ascertain the reliability of their predicted electronic excitation energies. Here, we have performed benchmark calculations aimed at assaying the M06-L, M06, M06-2X, and M06-HF functionals for TD-

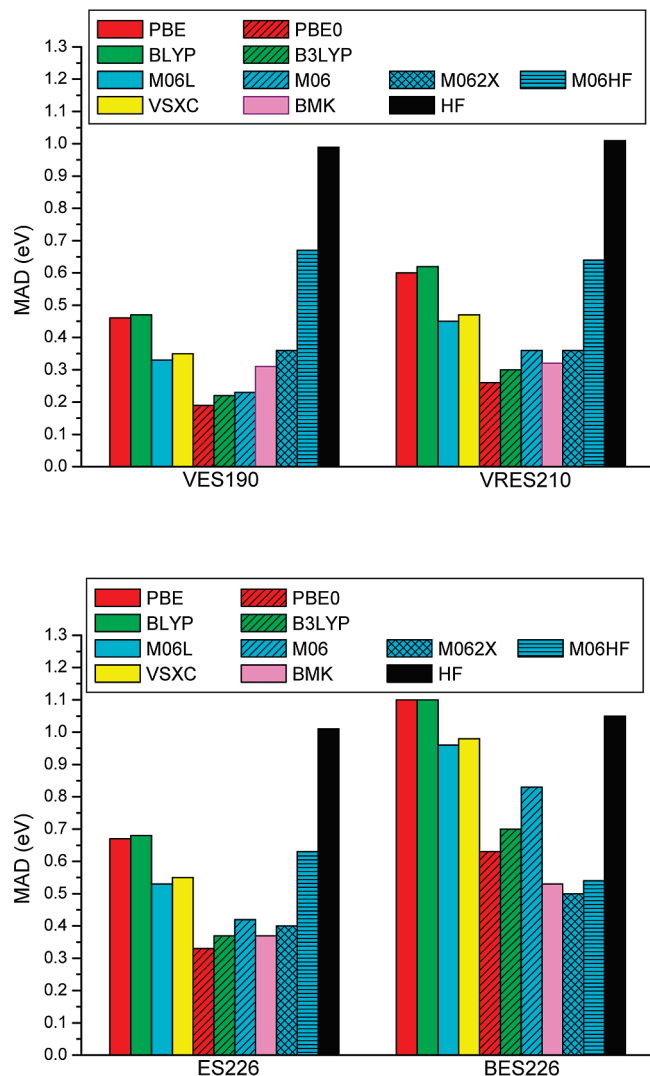


Figure 5. Computed MAD (eV) from the best estimates for combined data sets using different functionals.

DFT calculations. No functional, either one in the M06 family or any other considered functional, shows acceptable accuracy for all three of valence, Rydberg, and charge transfer excitations. Thus, we focus on valence excitations, for which the accuracy is most useful. The conclusions obtained through comparisons with theoretical benchmarks for isolated molecules and reference values inferred from experimental measurements in liquid solvents appear consistent and are as follows:

1. The M06 functional yields average deviations for valence excitations similar to those obtained with other popular hybrid functionals, namely, B3LYP and PBE0. The mean unsigned deviation (MAD) from the best estimates for the set of 190 valence excitations for which we have the most comprehensive group of comparisons set is 0.23 eV, but the deviations are significantly system-dependent. For example, the MAD is 0.39 eV for 18 $n \rightarrow \pi^*$ transitions of neutral nitroso dyes, but only 0.11 eV for a set of 49 $\pi \rightarrow \pi^*$ transitions of a variety of neutral dyes (see Tables 3–5). (In contrast, the MAD for excitations of charged cyanine dyes, which are not grouped with the valence excitations, is 0.55 eV.)

2. M06-L outperforms BP86 for the set of 103 valence-excitation benchmarks based on high-level wave function theory, and it leads to smaller errors than other meta-GGA functionals (VSXC, TPSS, and τ -HCTH) in the majority of cases. Although it tends, as do all local functionals, to underestimate the transition energies for many classes of excitation, this functional represents an improvement as compared to other local functionals.

3. Overall, M06-2X appears slightly less accurate than M06 for evaluating valence transition energies, although such a conclusion must be considered as tentative because of the lack of vibronic modeling in the majority of our VE calculations. For the VT set, the use of M06-2X improves significantly over the BHLYP estimates.

4. M06-HF is the least accurate among the four functionals of the M06 family for valence transitions. It significantly overestimates the transition energies for most $\pi \rightarrow \pi^*$ states, although the deviations are smaller than with the TD-HF approach.

In general, the trends and average errors found here for the four functionals of the M06 family are not inconsistent with expectations based on previous work.^{10,14,23}

Acknowledgment. D.J. and E.A.P. thank the Belgian National Fund for Scientific Research for their research associate and senior research associate positions, respectively. Several calculations have been performed on the Interuniversity Scientific Computing Facility (ISCF), installed at the Facultés Universitaires Notre-Dame de la Paix (Namur, Belgium), for which the authors gratefully acknowledge the financial support of the FNRS-FRFC and the “Loterie Nationale” for the convention number 2.4578.02 and of the FUNDP. The collaboration between the Belgian and French groups is supported by Wallonie-Bruxelles International, the Fonds de la Recherche Scientifique, the Ministère Français des Affaires Étrangères et Européennes, the Ministère de l’Enseignement Supérieur et de la Recherche, in the framework of Hubert Curien Partnership. The work at the University of Minnesota was supported in part by a grant to D.G.T. by the National Science Foundation (NSF). The visits of I.L. and C.A. to the University of Minnesota were supported in part by the NSF through a grant to the Institute for Mathematics and its Applications (IMA).

Supporting Information Available: Many additional details of the calculations and further discussion. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997.
- (2) Casida, M. E. In *Time-Dependent Density-Functional Response Theory for Molecules*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; Vol. 1, pp 155–192.
- (3) Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454.
- (4) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218.
- (5) Caricato, M.; Mennucci, B.; Tomasi, J. *J. Phys. Chem. A* **2004**, *108*, 6248.

- (6) Perdew, J. P.; Ruzsinsky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 62201.
- (7) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009.
- (8) Jacquemin, D.; Perpète, E. A. *Chem. Phys. Lett.* **2006**, *429*, 147.
- (9) Peach, M. J. G.; Cohen, A. J.; Tozer, D. *J. Phys. Chem. Chem. Phys.* **2006**, *8*, 4543.
- (10) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.
- (11) Barone, V.; Polimeno, A. *Chem. Soc. Rev.* **2007**, *36*, 1724.
- (12) Jacquemin, D.; Perpète, E. A.; Vydrov, O. A.; Scuseria, G. E.; Adamo, C. *J. Chem. Phys.* **2007**, *127*, 94102.
- (13) Ciofini, I.; Adamo, C. *J. Phys. Chem. A* **2007**, *111*, 5549.
- (14) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215. Unpublished errata: VES21, VRES41, CTE3, and ES44 should be VES20, VRES40, CTE3, and ES43. The final two Rydberg transitions were incorrectly labeled as valence in Table 14 of this reference, and one charge transfer transition was misassigned for M06-HF. These errors are all corrected in the present paper.
- (15) Matsuura, M.; Sato, H.; Sotoyama, W.; Takahashi, A.; Sakurai, M. *J. Mol. Struct. (THEOCHEM)* **2008**, *860*, 119.
- (16) Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *Chem. Phys. Lett.* **2008**, *465*, 226.
- (17) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 44118.
- (18) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.
- (19) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Acc. Chem. Res.* **2009**, *42*, 326.
- (20) Rohrdanz, M. A.; Martins, K. M.; Herbert, J. M. *J. Chem. Phys.* **2009**, *130*, 54112.
- (21) Goerigk, L.; Moellmann, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4611.
- (22) Fabian, J. *Dyes Pigm.* **2010**, *84*, 36.
- (23) Jacquemin, D.; Wathélet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420.
- (24) Caricato, M.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Theory Comput.* **2010**, *6*, 370.
- (25) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.
- (26) Cossi, M.; Barone, V. *J. Chem. Phys.* **2001**, *115*, 4708.
- (27) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.
- (28) Scalmani, G.; Frisch, M. J.; Mennucci, B.; Tomasi, J.; Cammi, R.; Barone, V. *J. Chem. Phys.* **2006**, *124*, 94107.
- (29) (a) Bondar, A.-N.; Fischer, S.; Smith, J.; Elstner, M.; Suhai, S. *J. Am. Chem. Soc.* **2004**, *126*, 14668. (b) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458.
- (30) Curutchet, C.; Scholes, G. D.; Mennucci, B.; Cammi, R. *J. Phys. Chem. B* **2007**, *111*, 13253.
- (31) Jacquemin, D.; Perpète, E. A.; Laurent, A. D.; Assfeld, X.; Adamo, C. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1258.
- (32) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.
- (33) Romaniello, P.; Sangalli, D.; Berger, J. A.; Sottile, F.; Molinari, L. G.; Reining, L.; Onida, G. *J. Chem. Phys.* **2009**, *130*, 44108.
- (34) Gritsenko, O. V.; Baerends, E. J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4640.
- (35) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.
- (36) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
- (37) Świderek, K.; Paneth, P. *J. Phys. Org. Chem.* **2009**, *22*, 845.
- (38) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 1095.
- (39) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849.
- (40) Korth, M.; Grimme, S. *J. Chem. Theory Comput.* **2009**, *5*, 993.
- (41) Yang, K.; Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2010**, *132*, 164117.
- (42) Averkiev, B. B.; Zhao, Y.; Truhlar, D. G. *J. Mol. Catal. A* **2010**, DOI: 10.1016/j.molcata.2010.03.016.
- (43) Zheng, J. J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 569.
- (44) Zheng, J. J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 808.
- (45) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. C* **2008**, *112*, 6860.
- (46) Valero, R.; Gomes, J. R. B.; Truhlar, D. G.; Illas, F. *J. Chem. Phys.* **2008**, *129*, 124710.
- (47) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 324.
- (48) Valero, R.; Gomes, J. R. B.; Truhlar, D. G.; Illas, F. *J. Chem. Phys.* **2010**, *132*, 104701.
- (49) Sorkin, A.; Iron, M. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 307.
- (50) Dahlke, E. E.; Olson, R. M.; Leverentz, H. R.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 3976.
- (51) Leverentz, H. R.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 6009.
- (52) (a) Sorkin, A.; Truhlar, D. G.; Amin, E. A. *J. Chem. Theory Comput.* **2009**, *5*, 1254. (b) Mantina, M.; Valero, R.; Truhlar, D. G. *J. Chem. Phys.* **2009**, *131*, 64706. (c) Ferrighi, L.; Hammer, B.; Madsen, G. K. H. *J. Am. Chem. Soc.* **2009**, *131*, 10605.
- (53) Liao, M. S.; Watts, J. D.; Huang, M. J. *J. Chem. Theory Comput.* **2008**, *7*, 615.
- (54) Jimenez-Hoyos, C. A.; Janesko, B. G.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6621.
- (55) Biczysko, M.; Panek, P.; Barone, V. *Chem. Phys. Lett.* **2009**, *475*, 105.
- (56) Valdes, H.; Pluhackova, K.; Pitonak, M.; Rezac, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2747.
- (57) (a) van Mourik, T. *J. Chem. Theory Comput.* **2008**, *4*, 1610. (b) Cao, J.; van Mourik, T. *Chem. Phys. Lett.* **2010**, *485*, 40.
- (58) Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. *J. Chem. Theory Comput.* **2008**, *4*, 1996.
- (59) Dahlke, E. E.; Olson, R. M.; Leverentz, H. R.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 3976.
- (60) Bryantsev, V. S.; Diallo, M. S.; van Duin, A. C. T.; Goddard, W. A. *J. Chem. Theory Comput.* **2009**, *5*, 1016.

- (61) Raju, R. K.; Ramarj, A.; Hillier, I. H.; Vincent, M. A.; Burton, N. A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 3411.
- (62) Valero, R.; Costa, R.; Moreira, I. D. P. R.; Truhlar, D. G.; Illas, F. *J. Chem. Phys.* **2008**, *128*, 114103.
- (63) Ruiz, E. *Chem. Phys. Lett.* **2008**, *460*, 336.
- (64) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, *112*, 6794.
- (65) Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 2284.
- (66) Frisch, M. J. *Gaussian 03*, revisions D.02 and E.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (67) Frisch, M. J. *Gaussian DV*, revision H.01; Gaussian, Inc.: Wallingford, CT, 2008.
- (68) Zhao, Y.; Truhlar, D. G. *MN-GFM*, version 4.1; University of Minnesota: Minneapolis, MN, 2008.
- (69) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Gaussian, Inc., Wallingford CT, 2009.
- (70) Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123.
- (71) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.
- (72) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029.
- (73) Adamo, C.; Scuseria, G. E.; Barone, V. *J. Chem. Phys.* **1999**, *111*, 2889.
- (74) Rohrdanz, M. A.; Herbert, J. M. *J. Chem. Phys.* **2008**, *129*, 34107.
- (75) Impropa, R.; Barone, V.; Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2006**, *125*, 54103.
- (76) Li, J.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **2000**, *77*, 264.
- (77) Hammick, D. L.; Lister, M. W. *J. Chem. Soc.* **1937**, 489.
- (78) Schroeder, W. A.; Wilcox, P. E.; Trueblood, K. N.; Dekker, A. O. *Anal. Chem.* **1951**, *23*, 1740.
- (79) Haszeldine, R. N.; Jander, J. *J. Chem. Soc.* **1954**, 691.
- (80) Tarte, P. *Bull. Soc. Chim. Belg.* **1954**, *63*, 525.
- (81) Jander, J.; Haszeldine, R. N. *J. Chem. Soc.* **1954**, 912.
- (82) Bugai, P. M.; Konel'skaya, V. N. *Izvest. Akad. Nauk. SSSR - Ser. Fizich.* **1954**, *18*, 695.
- (83) Labhart, H. *Helv. Chim. Acta* **1957**, *152*, 1410.
- (84) Mason, J. *J. Chem. Soc.* **1957**, 3904.
- (85) Mason, J. *J. Chem. Soc.* **1959**, 1288.
- (86) Lutskii, A. E.; Konel'skaya, V. N. *Z. Obs. Khim.* **1960**, *30*, 3773.
- (87) Bugai, P. M.; Konel'skaya, V. N.; Gol'berkova, A. S.; Bazhenova, L. M. *Z. Fizich. Khim.* **1962**, *36*, 2233.
- (88) Tabei, K.; Nagakura, S. *Bull. Soc. Chim. Japan* **1965**, *38*, 965.
- (89) Asquith, R. S.; Bridgeman, I.; Peters, A. T. *J. Soc. Dyers Colour.* **1965**, *81*, 439.
- (90) Bell, M. G. W.; Day, M.; Peters, A. T. *J. Soc. Dyers Colour.* **1966**, *82*, 410.
- (91) Asquith, R. S.; Peters, A. T.; Wallace, F. *J. Soc. Dyers Colour.* **1968**, *84*, 507.
- (92) Day, M.; Peters, A. T. *J. Soc. Dyers Colour.* **1969**, *85*, 8.
- (93) Mason, J. *J. Chem. Soc. A* **1969**, 1587.
- (94) Matsubayashi, G. E.; Takaya, Y.; Tanaka, T. *Spectrochim. Acta* **1970**, *26A*, 1851.
- (95) Weast, R. C. *Handbook of Chemistry and Physics*, 51st ed.; The Chemical Rubber Company: Cleveland, OH, 1970; p 1.
- (96) Grasselli, J. G. *Atlas of Spectral Data and Physical Constants for Organic Compounds*; The Chemical Rubber Company: Cleveland, OH, 1973; p 1.
- (97) Thomson, R. H. *Naturally Occurring Quinones*, 2nd ed.; Academic Press: London, 1971; pp 1.
- (98) Allston, T. D.; Fedyk, M. L.; Takacs, G. A. *Chem. Phys. Lett.* **1978**, *60*, 97.
- (99) Fabian, J.; Hartmann, H. *Light Absorption of Organic Colorants*; Vol. 12; Springer-Verlag: Berlin, 1980; pp 1.
- (100) Fabian, J.; Nepras, M. *Collect. Czech. Chem. Commun.* **1980**, *45*, 2605.
- (101) Csaszar, J. *Acta Phys. Chem. (Szeged)* **1987**, *33*, 11.
- (102) Green, F. J. *The Sigma-Aldrich Handbook of Stains, Dyes, and Indicators*; Aldrich Chemical Company: Milwaukee, WI, 1990; p 1.
- (103) Alexiou, M. S.; Tychopoulos, V.; Ghorbanian, S.; Tyman, J. H. P.; Brown, R. G.; Brittain, P. I. *J. Chem. Soc. Perkin Trans. 2* **1990**, 837.
- (104) Wintgens, V.; Valat, P.; Kossanyi, J.; Biczok, L.; Demeter, A.; Berces, T. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 411.
- (105) Biczok, L.; Valat, P.; Wintgens, V. *Phys. Chem. Chem. Phys.* **1999**, *1*, 4759.
- (106) Günaydin, K.; Topcu, G.; Ion, R. M. *Nat. Prod. Lett.* **2002**, *16*, 65.
- (107) (a) Parac, M.; Grimme, S. *J. Phys. Chem. A* **2002**, *106*, 6844. (b) Diercksen, M.; Grimme, S. *J. Chem. Phys.* **2004**, *120*, 3544. (c) Santoro, F.; Impropa, R.; Lami, A.; Bloino, J.; Barone, V. *J. Chem. Phys.* **2007**, *126*, 84509.
- (108) Renge, I. *J. Phys. Chem. A* **2009**, *113*, 10678–10686.
- (109) Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69.
- (110) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (111) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (112) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (113) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648. (b) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

- (114) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.
- (115) (a) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540. (b) Rohrdanz, M. A.; Herbert, J. M. *J. Chem. Phys.* **2008**, *129*, 34107.
- (116) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.
- (117) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51.
- (118) Perdew, J. P.; Ruzsinszky, A.; Constantin, L. A.; Sun, J.; Csonka, G. I. *J. Chem. Theory Comput.* **2009**, *5*, 902.
- (119) Grimme, S.; Waletzke, M. *J. Chem. Phys.* **1999**, *111*, 5645.
- (120) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4786.
- (121) Griffiths, J. *Colour and Constitution of Organic Molecules*; Academic Press: London, 1976; p 1.
- (122) Christie, R. M. *Colour Chemistry*; The Royal Society of Chemistry: Cambridge, UK, 1991; p 228.
- (123) Zollinger, H. *Color Chemistry, Syntheses, Properties and Applications of Organic Dyes and Pigments*, 3rd ed.; Wiley-VCH: Weinheim, Germany, 2003; p 647.
- (124) Jacquemin, D.; Preat, J.; Charlot, M.; Wathélet, V.; André, J. M.; Perpète, E. A. *J. Chem. Phys.* **2004**, *121*, 1736.
- (125) Jacquemin, D.; Preat, J.; Wathélet, V.; André, J. M.; Perpète, E. A. *Chem. Phys. Lett.* **2005**, *405*, 429.
- (126) Perpète, E. A.; Wathélet, V.; Preat, J.; Lambert, C.; Jacquemin, D. *J. Chem. Theory Comput.* **2006**, *2*, 434.
- (127) Jacquemin, D.; Assfeld, X.; Preat, J.; Perpète, E. A. *Mol. Phys.* **2007**, *105*, 325.
- (128) Jacquemin, D.; Wathélet, V.; Preat, J.; Perpète, E. A. *Spectrochim. Acta A* **2007**, *67*, 334.
- (129) Preat, J.; Laurent, A. D.; Michaux, C.; Perpète, E. A.; Jacquemin, D. *J. Mol. Struct. (THEOCHEM)* **2009**, *901*, 24.
- (130) Jacquemin, D.; Perpète, E. A.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105.
- (131) Jacquemin, D.; Bouhy, M.; Perpète, E. A. *J. Chem. Phys.* **2006**, *124*, 204321.
- (132) Demeter, A.; Berces, T.; Biczok, L.; Wintgens, V.; Valat, P.; Kossanyi, J. *J. Phys. Chem.* **1996**, *100*, 2001.
- (133) Jacquemin, D.; Perpète, E. A.; Scalmani, G.; Frisch, M. J.; Ciofini, I.; Adamo, C. *Chem. Phys. Lett.* **2007**, *448*, 3.
- (134) Miao, L.; Yao, Y.; Yang, F.; Wang, Z.; Li, W.; Hu, J. *J. Mol. Struct. (THEOCHEM)* **2008**, *865*, 79.
- (135) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.
- (136) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
- (137) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.
- (138) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, *108*, 9624.
- (139) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129.
- (140) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.
- (141) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (142) Jacquemin, D.; Perpète, E. A. *Chem. Phys. Lett.* **2006**, *420*, 529.
- (143) Schreiber, M.; Bub, V.; Fülischer, M. P. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3906.
- (144) Guillaumont, D.; Nakamura, S. *Dyes Pigm.* **2000**, *46*, 85.
- (145) Fabian, J. *Theor. Chem. Acc.* **2001**, *106*, 199.
- (146) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116.
- (147) Buenker, R. J.; Peyerimhoff, S. D.; Kammer, W. D. *J. Chem. Phys.* **1971**, *555*, 814.
- (148) Buenker, R. J.; Peyerimhoff, S. D. *Chem. Phys.* **1975**, *36*, 415.
- (149) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425.
- (150) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943.
- (151) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.
- (152) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
- (153) Xu, X.; Goddard, W. A., III *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673.
- (154) Serrano-Andrés, L.; Merchán, M.; Nebot-Gil, I.; Lindh, R.; Roos, B. O. *J. Chem. Phys.* **1993**, *98*, 3151.
- (155) (a) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403. (b) Baker, J.; Pulay, P. *J. Chem. Phys.* **2002**, *117*, 1441.
- (156) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559.
- (157) Tao, J.; Perdew, J.; Staroverov, V.; Scuseria, G. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (158) (a) Slater, J. C. *Phys. Rev.* **1951**, *81*, 385. (b) Vosko, S. J.; Wilk, L.; Nusair, M. *Can. J. Phys. Can. J. Phys.* **1980**, *58*, 1200.
- (159) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425.

CT100119E

Can Range-Separated and Hybrid DFT Functionals Predict Low-Lying Excitations? A Tookad Case Study

Boxue Tian,[†] Emma S. E. Eriksson,[†] and Leif A. Eriksson^{*,†,‡}

School of Chemistry, National University of Ireland - Galway, Galway, Ireland and Örebro Life Science Center, School of Science and Technology, Örebro University, 701 82 Örebro, Sweden

Received March 18, 2010

Abstract: The spectral properties of Tookad (Pd-bacteriopheophorbide, Pd-BPheid), an effective photosensitizer that targets mainly prostate tumors, and metal-free BPheid have been studied using time-dependent density functional theory (TD-DFT). The well-established B3LYP functional, which is known to overestimate excitation energies, was included in the study along with recently introduced range-separated and meta hybrid DFT functionals CAM-B3LYP, M06, M06-2X, M06HF, ω B97XD, ω B97X, ω B97, LC- ω PBE, and PBE0 (PBE1PBE). The main focus is the performance of the new functionals in predicting low-lying excitations (>600 nm), to explore their potential roles in drug development for photodynamic therapy. The data suggests that ω B97XD overall performs best for the Q_y transition band (the red-most absorption), followed by CAM-B3LYP. LC- ω PBE, ω B97, B3LYP, and PBE1PBE all generated the Q_y band far from the experimental position. The error in absorption energy for the Q_y band was found to be at most 0.05 eV for ω B97XD, compared to 0.15–0.19 eV for B3LYP. The use of different basis sets used in excited-state calculations was shown to be of less importance as was the use of either B3LYP or M06 in geometry optimizations.

1. Introduction

Photodynamic therapy (PDT), in which a photosensitizer, light, and oxygen are the major components, has been shown to be a promising method for treatment of various cancers as well as other diseases. In the reaction between the excited photosensitizer and oxygen in the tissue, reactive oxygen species (ROS), such as singlet oxygen, are formed and can readily react with the target tissue. Photosensitizers are light-absorbing molecules often made up by conjugated systems, such as porphyrins, chlorins (17,18-dihydroporphyrin), and bacteriochlorins (7,8,17,18-tetrahydroporphyrin). The first approved and most widely used photosensitizer is Photofrin that has been successfully used in PDT. However, Photofrin suffers from drawbacks, such as light absorption at wavelengths below the optimal tissue penetration as well as long-lasting photosensitization, due to accumulation in the skin

tissue (low-clearance rate). Additional photosensitizers are now available on the market, and new photosensitizers are continuously being developed with the aim to reduce the side effects and increase the efficiency of the treatment.

One of the most important aspects in the development of photosensitizers is the absorption properties. The red-most absorption peak of porphyrin- and chlorin-based photosensitizers is in general positioned between 600 and 700 nm and is the one used in PDT to excite the photosensitizer. Although this is usually significantly weaker than absorptions occurring around 400 nm, it is used in PDT due to the increased tissue penetration of the light at these wavelengths. Bacteriochlorophylls (BChl) display relatively speaking very large extinction coefficients for the low-lying Q_y band, and thus some BChls and their derivatives have been suggested to be utilized as photosensitizers in PDT.^{1,2} Substitution of the central Mg^{2+} in native BChl with other divalent transition-metal ions, such as Pd^{2+} , Co^{2+} , Ni^{2+} , Cu^{2+} , Zn^{2+} , and Mn^{2+} has been carried out successfully,³ and the spectroscopic properties of these synthesized compounds

* Corresponding author. E-mail: leif.eriksson@nuigalway.ie.

[†] National University of Ireland.

[‡] Örebro University.

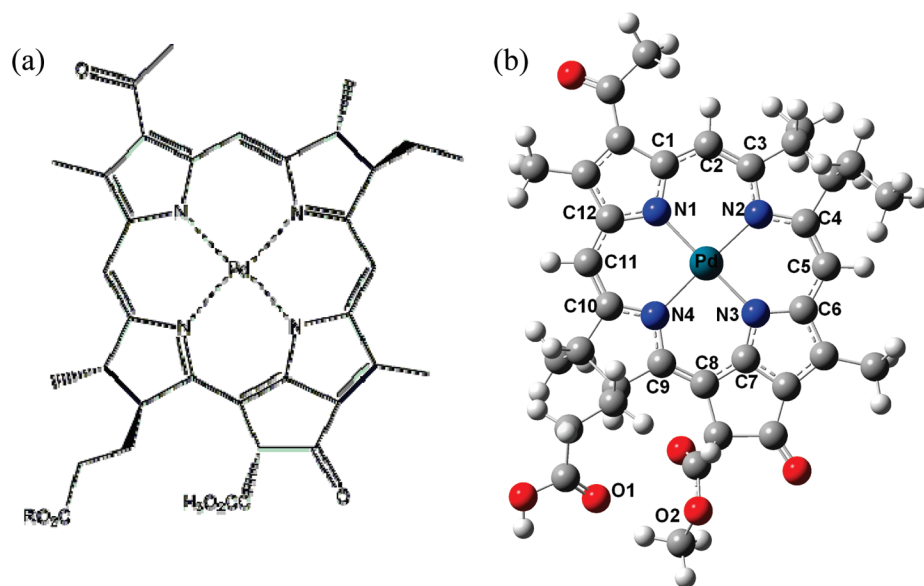


Figure 1. (a) Schematic figure of the Tookad molecule. R = H: Tookad (Pd-BPheid); R = C₂₀H₃₉: Pd-BChl. (b) Optimized structure of the Tookad molecule.

have been extensively studied. BChl with the central metal removed and replaced by two hydrogen atoms (bacteriopheophytin, BPhe) displays its two red-most absorption peaks at 749–758 and at 524–531 nm in diethyl ether (DE),^{3–5} toluene, and tetrahydrofuran (THF),⁴ corresponding to the Q_y and Q_x transition bands, respectively. Metal-containing BChls absorb strongly at wavelengths in the range of 753–782 nm, and the second absorption peak is observed at 529–594 nm in DE and THF.^{3,6} From these data, it is clear that the change of metal in the molecular center significantly affects the position of the Q_x band, whereas the position of the low-lying Q_y band is less dependent on the metal.

Pd-containing BChl derivatives have achieved particular attention due to enhanced stability and promising photodynamic properties. Tookad (WST09), displayed in Figure 1, is a Pd-containing BChl derivative in which the phytyl group (C₂₀H₃₉) at the propionyl residue has been replaced by a hydrogen atom, giving Pd-bacteriopheophorbide (Pd-BPheid). However, the presence or absence of the phytyl group does not affect the position of the absorption bands as the phytyl group does not contribute to the conjugation,^{7,8} and spectroscopic data reported for Pd-BChl and Tookad are therefore almost identical. Pd-BChl displays the Q_y and Q_x absorption bands at 753–762 nm and 527–535 nm in DE,^{3,5} toluene,⁹ and THF.⁶ For Pd-BChl and Tookad, the quantum yield for intersystem crossing from the first-excited singlet state to the triplet state is nearly 100%, which results in a large number of triplet-state molecules available to react with molecular oxygen, with the main product being singlet oxygen.^{9,10}

In vitro and in vivo studies on the effect of Tookad-PDT have showed significant phototoxicity on several different carcinoma cells and tumors as a result of tumor vascular damage.^{11–17} The promising results from these studies have led to phase I^{18–20} and II²¹ clinical trials performed on patients with prostate cancer. Tookad has, as opposed to several other photosensitizers, a fast clearance rate from the

body.²² It does not seem to accumulate in skin, muscle, and tumor tissue but mainly in the plasma, kidney, and liver, from which it is cleared relatively rapidly.²² The low accumulation in skin tissue would reduce the long-lasting photosensitization side effects common for photosensitizers, such as Photofrin.

Computational modeling holds good promise for drug development and refinement, in that we are able to process large quantities of data in a short period of time and thus point to modifications (or entirely new molecules) that would be of interest to synthesize and assess experimentally. In the current context, if we are to find improved chromophores for PDT, then the capability to compute accurate absorption spectra would be essential for the methodology chosen. As the molecules discussed herein are rather large, and it would be desirable to explore changes in spectra for a significant number of possible substitutions, excited-state calculations within the time-dependent density functional theory (TD-DFT) formalism is an attractive option.

TD-DFT has been employed successfully over the past decade to explore both spectra and photochemical properties on a wide range of systems. However, an early comparative study of the performance of TD-DFT versus CASSCF/CASPT2 (multiconfigurational SCF and second-order perturbation theory) on excited-state calculations of a number of organic compounds showed that the methodology suffers from defects that makes it less accurate compared with pure ab initio methods.²³ Several benchmarking studies of excited-state calculations using more recent DFT functionals have since been reported. Perpète et al used B3LYP and PBE0 to explore absorptions of 66 different substituted anthraquinones, showing that PBE0 was able to provide data to within a mean average deviation (MAD) of 0.05 eV after the application of a fitting procedure for this particular set of systems.²⁴ Andzelm et al used eight different functionals (including HF and local density functionals) for a set of tricyanofuran based push–pull (donor–acceptor) chromophores.²⁵ Again the PBE0 functional, along with CAM-

B3LYP and BNL, were found to perform the best after adjustment of the attenuation factor γ . However, the data showed large spread, and issues, such as transferrability and uneven performance for charge-transfer versus π - π^* excitations, remain to be resolved. Two larger functional benchmark studies have been reported by Jacquemin et al, dealing with singlet excited states and singlet-triplet gaps, respectively.^{26,27} In their extensive study of singlet excitations, 29 functionals were tested, computing 700 excitations for 500 organic molecules of varying sizes.²⁶ The data show a very large system dependence, with the best MAD of about 0.25 eV. Functionals containing a large fraction of exchange significantly underestimated the excitation energies, and overall the functionals of 'standard hybrid' type, such as X3LYP, B3LYP and PBE0, performed the best. In the study of singlet-triplet gaps, finally, 34 functionals were included to study a total of 63 excited states in 20 medium-sized molecules.²⁷ Again, the functionals displayed a large spread, with functionals both over- and under-shooting; the MAD varying between 0.2–0.7 eV. The BMK and M06-2X functionals were in this case found to perform the best, albeit M06-2X is unpredictable in that it sometimes gives too high and sometimes too low values; PBE0 and CAM-B3LYP were also among the better-performing functionals but consistently overshoot the experimental excitation energies.

A large number of computational studies on photosensitizing compounds based on porphyrin, chlorin, and bacteriochlorin structures have also been reported. As in the benchmarking examples outlined above, excited-state calculations have been performed using different methods, generating deviating results. Palma et al. recently summarized the present computational results on free-base porphyrin and chlorin obtained by different methods and functionals.²⁸ It is clear from these data that TD-DFT in combination with any functional used in those studies overestimates the excitation energy of the Q_y band by 0.11–0.34 eV. Excited-state calculations of metal-coordinated porphyrins, e.g. Zn-porphyrin,^{29,30} performed at the TD-DFT/B3LYP level of theory, also overestimated the excitation energies.

A four-orbital model has been suggested in order to explain the four transitions (two Q and two B bands) seen in the absorption spectra of porphyrins.³¹ This model only considers the two highest occupied molecular orbitals (HOMO-1 and HOMO) and the two lowest unoccupied molecular orbitals (LUMO and LUMO+1) and the four possible excitations between those. However, the four-orbital model has been questioned as both experimentally and theoretically more than four transition bands have been found. Still the question is being discussed as diverging results are being generated, and different conclusions are drawn. In the case of more than four transition bands found the result can be interpreted either as if the additional bands found represent vibrational overtones of the electronic transition bands,³² hence the four-orbital theory would hold, or as if all observed bands are electronic transitions,³³ meaning that the four-orbital model would be inappropriate. Recent theoretical studies performed on chlorophyll a and pheophytin a have not thrown further light on the issue. Symmetry-adapted cluster configuration interaction (SAC-CI) calculations³⁴ support the four-orbital

model by generating only four transition bands, in agreement with the ones found by Houssier et al.³² At the TD-DFT/Becke-Perdew (TD-DFT/B-P)^{7,8} and DFT with multireference configuration interaction (DFT/MRCI)³⁵ level of theory however, apart from the four transition bands, additional bands are found in-between the Q bands as well as at higher energies, results that support the assumption that the four-orbital model is too simple to correctly describe the absorption spectra.³³ From these studies it is also concluded that TD-DFT overestimates the excitation energy of the Q_y band.

In order to try to establish a suitable methodology for the study of these types of systems, we have in the present study investigated Tookad and metal-free BPheid with the aim to assess the performance of nine recent functionals in predicting low-lying excited-state energies, i.e., evaluating their predictive power and hence potential to use in computer-assisted drug design for PDT. The main question to be answered is if either of these functionals can reproduce the long wavelength peak of the absorption spectrum, a task in which several commonly applied functionals today fail. We used B3LYP,³⁶ PBE0³⁷ (PBE1PBE), LC- ω PBE,^{38–41} CAM-B3LYP,⁴² ω B97,⁴³ ω B97X,⁴³ ω B97XD⁴⁴, M06,⁴⁵ M06-2X,⁴⁵ and M06HF^{46,47} for excited-state calculations and B3LYP and M06 for geometry optimizations. In order to explore if the presence or absence of a metal in the compound affects the performance of the functionals, both Tookad (Pd-BPheid) and metal-free BPheid were included.

LC- ω PBE, CAM-B3LYP, and ω B97XD are so-called long-range corrected (LC) functionals in which the Coulomb r_{12}^{-1} term is split into a long-range part that includes the Hartree-Fock (HF) exchange integral and a short-range part that includes the DFT exchange interaction. The ω parameter is introduced to control the range of the interelectronic separation between the two terms. LC- ω PBE is the long-range corrected version of the nonempirical generalized gradient approximation (GGA) functional ω PBE (Perdew-Burke-Ernzerhof). CAM-B3LYP uses a Coulomb attenuating method to combine the hybrid B3LYP method with a long-range correction by introducing two extra parameters, instead of the single parameter used by Becke. ω B97XD is an extended version of the long-range corrected ω B97 and ω B97X functionals with empirical dispersion correction. M06HF is a full-HF exchange functional with satisfying long-range properties, primarily designed for Rydberg and charge-transfer excitations. M06 and M06-2X, finally, are extensions of M06HF, with focus on valence excitations.

2. Computational Methodology

All calculations were carried out using the Gaussian09 program.⁴⁸ Neutral Tookad and the corresponding metal-free BPheid were studied. The geometries of the ground singlet state of the molecules were initially optimized in vacuum, toluene, and THF using B3LYP and M06. The LanL2DZ⁴⁹ basis set, an effective core potential (ECP) and valence electron (double- ζ) basis set combination, was used for Pd, and the 6-31+G(d,p) all-electron basis set was used for all other atoms in the geometry optimizations. The bulk solvation was modeled through the integral equation formalism of the polarized continuum model (IEF-PCM).^{50,51}

Table 1. Selected Geometrical Parameters for Tookad Optimized in Toluene with B3LYP and M06 in Conjunction with LanL2DZ Basis Set for Pd and 6-31+G(d,p) for All Other Atoms^a

atoms	distance (Å)	
	B3LYP	M06
Pd–N1	2.054	2.048
Pd–N2	2.062	2.050
Pd–N3	2.015	2.010
Pd–N4	2.114	2.100
(N1···N3)	4.069	4.058
(N2···N4)	4.175	4.149
N1–C1	1.367	1.362
C1–C2	1.400	1.396
C2–C3	1.395	1.381
C3–N2	1.357	1.351
N2–C4	1.379	1.376
C4–C5	1.377	1.373
C5–C6	1.406	1.403
C6–N3	1.384	1.378
N3–C7	1.337	1.332
C7–C8	1.399	1.395
C8–C9	1.377	1.372
C9–N4	1.381	1.375
N4–C10	1.359	1.353
C10–C11	1.384	1.381
C11–C12	1.398	1.393
C12–N1	1.371	1.366
(O1···O2)	3.499	3.142

^a Atom names correspond to the labels in Figure 1b.

Vertical singlet excitation energies were calculated in vacuum, toluene, and THF for Tookad and BPheid using both structures optimized with B3LYP and M06, employing the time-dependent (TD) formalism^{52–54} and the B3LYP, M06HF, and ω B97XD functionals. For the Tookad and BPheid geometries optimized using B3LYP, vertical singlet excitation energies in toluene were also calculated using CAM-B3LYP, M06, M06-2X, ω B97X, ω B97, LC- ω PBE, and PBE0. The LanL2DZ and 6-311+G(2d,2p) basis sets were used in the excited-state calculations. The basis sets 6-31G(d,p) and 6-31+G(d,p) (for all atoms but Pd) were also included in a test set of excited-state calculations with ω B97XD on the Tookad B3LYP geometry in toluene. The calculated wavelengths (in nm) are plotted against the oscillator strengths using a Gaussian line shape.

3. Results

3.1. Geometry. First the effect on the optimized geometries by the two different functionals was investigated. This is an important aspect to consider as geometrical changes can greatly influence the excitation properties of the molecule. Selected geometrical parameters for Tookad optimized in toluene using B3LYP and M06 are shown in Table 1, and Cartesian coordinates are provided in the Supporting Information. The data show that M06 generates a structure with an overall more ‘compact’ conjugated core.

The tail parts of the molecule (represented by the O1···O2 distance in Table 1 and Figure 1) are most affected by the choice of functional and can, without explicitly taking part in the conjugation, influence the conjugated system by affecting the geometry of the ring system. However, the small differences in lengths for the bonds participating in the

Table 2. Calculated Low-Lying Absorption Bands (Q_y and Q_x) for Tookad (Pd-BPheid) and BPheid in Toluene and THF^a

geometry	TD-DFT	solvent = toluene		solvent = THF			
		Q_y	Q_x	Q_y	Q_x		
Tookad							
		Exptl ⁹	762	535	Exptl ⁶	755	529
B3LYP	B3LYP		695	528		693	530
	M06HF		810	523		743	521
	ω B97XD		769	528		750	527
M06	B3LYP		690	524		688	526
	M06HF		805	519		745	517
	ω B97XD		763	523		746	522
BPheid							
		Exptl ⁴	758	531	Exptl ⁴	751	527
B3LYP	B3LYP		681	554		679	554
	M06HF		777	530		708	524
	ω B97XD		759	553		738	550
M06	B3LYP		679	549		678	549
	M06HF		760	524		689	517
	ω B97XD		753	547		729	544

^a Compared with experimental data for Pd-BChl and BPhe.

conjugated system have only a minor effect on the calculated spectra, as shown below.

The effect of the environment (gas phase, toluene, and THF) on the geometries is also mainly seen in the two tails of the molecule (data shown for toluene only, Table 1). The metal has a small effect on the geometry of the molecule. The presence of Pd in the structure attracts the nitrogen atoms in a way that the inner core of the ring system becomes more compact (atom distances N1···N3 and N2···N4 are reduced) compared to the metal-free system (data not shown).

3.2. Spectroscopic Properties. Table 2 shows the calculated lowest-lying absorption bands for Tookad and BPheid in toluene and THF, along with the experimental data for Pd-BChl and BPhe. As previously mentioned, the spectral properties should not be affected to any significant degree by the presence or absence of the phytyl group. The functionals B3LYP and M06 were used together with the LanL2DZ and 6-31+G(d,p) basis sets in the geometry optimizations, and B3LYP, M06HF, and ω B97XD in conjunction with the LanL2DZ and 6-311+G(2d,2p) basis sets were used in the excited-state calculations. Only the two red-most absorption bands (Q_x and Q_y) are discussed here, as the complete experimental spectra are not available for the compounds in both solutions. In addition, the Q_x and Q_y bands correspond to the absorption wavelengths of interest when studying properties related to PDT.

Excited-state calculations using the same functional (B3LYP, M06HF, or ω B97XD) on the B3LYP or M06 geometries generate absorption bands in the same range, within 2–6 and 1–19 nm for Tookad and BPheid, respectively. The M06 geometry results overall in slightly shorter wavelengths compared with the B3LYP geometry, however the difference is not significant.

The choice of functional used for excited-state calculations is more crucial, and the different functionals generate quite

widespread results. Excited-state calculations using B3LYP overall generates the Q_y band at shorter wavelengths than ω B97XD, whereas M06HF generates the Q_y band at longer or shorter wavelengths, depending on solvent. The choice of functional thus significantly affects the position of Q_y band, whereas the Q_x band is almost independent of the functional used. It is well-known that B3LYP in general overestimates excitation energies by ~ 0.1 – 0.2 eV, however when there is charge transfer involved the error should be larger. It is clear also from the results presented herein that the energies of the first excitation are overestimated using B3LYP. For Tookad, the excitation energy of the Q_y band is overestimated by 0.15 – 0.17 eV, whereas for BPheid the energy difference is slightly larger, 0.18 – 0.19 eV. For M06HF and ω B97XD, the results are inconsistent; in some cases, the excitation energies are either underestimated or overestimated, but overall, the results are closer to the experimental data than B3LYP. For M06HF, the error is in the range of 0.02 – 0.10 and 0 – 0.15 eV for Tookad and BPheid, respectively, and for ω B97XD, the error range is 0 – 0.02 and 0 – 0.05 eV for Tookad and BPheid, respectively. The error of the Q_x band is overall significantly smaller than for the Q_y band. It cannot be concluded from the generated results if the presence of the metal significantly influences the performance of the functionals. The solvent only affects the absorption peaks a few nm for B3LYP and ω B97XD, which is consistent with experimental data, whereas for M06HF the difference between the two solvents is significant for the Q_y band.

For Tookad, it is obvious that ω B97XD used in excited-state calculations generates absorption wavelengths closest to the experimental ones for the Q_y band. For the Q_x band, the performance of the functionals is more inconsistent; however, the difference between the functionals is small. ω B97XD and B3LYP generate results closest to experimental data in this case. For BPheid, ω B97XD again generates the best result for the Q_y band in three cases out of four. For the Q_x band, however, the functionals do not perform as equally as for Tookad, and the results show that M06HF generates the best result. However, for the Q_y band of both Tookad and BPheid, M06HF generates results far from experimental data in several cases. Gas-phase data are included in the Supporting Information. The absorption bands generated with B3LYP and ω B97XD in the gas phase are all blue-shifted compared with in bulk solvation. For M06HF, however, the Q_y band is in some cases not affected at all by the inclusion of bulk solvation, and the gas-phase data is also sometimes red-shifted compared with the data in bulk solvation. The Q_y band is most affected by the inclusion of bulk solvation, whereas the Q_x band is only red-shifted a few nm. These results confirm that including implicit solvent is in general necessary in order to obtain reasonable data.

In order to display the effect of the Pd atom on the absorption spectra, a comparison between Tookad and metal-free BPheid is displayed in Figure 2. Here we display the case in which ω B97XD was used for excited-state calculations in toluene on the B3LYP geometries. BPheid displays a more compressed spectrum, with the Q_y band blue-shifted, i.e., more energy is needed to excite BPheid to its first excited singlet state, compared with Tookad. The high-energy region

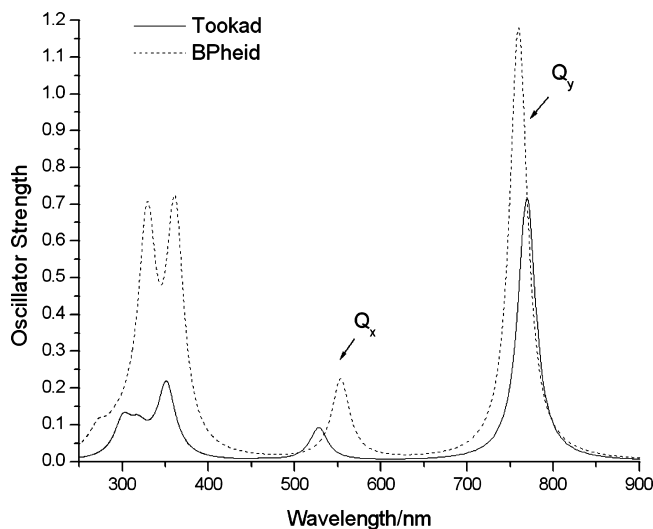


Figure 2. Absorption spectra of Tookad and BPheid generated using ω B97XD.

of the spectra is similar for the two molecules, however the oscillator strength for Tookad is much lower compared with that of BPheid. Also for the low-lying excitations, Tookad provides lower oscillator strengths. This finding is partly supported by experiments as Pd-BChl displays a lower extinction coefficient for the higher energy bands compared to BPhe, albeit a higher extinction coefficient for the Q_y band.³

For comparison purposes, seven additional functionals, CAM-B3LYP, M06, M062X, ω B97X, ω B97, LC- ω PBE, and PBE0 (PBE1PBE) were included in the calculation of the excited states of Tookad and BPheid in toluene using the B3LYP geometry. The resulting spectra are shown in Figures 3 and 4, together with the previously obtained data for B3LYP, M06HF, and ω B97XD. The spectra display a significant difference between the functionals, especially for the low-lying excitations. The Q_y band is clearly the strongest one, a feature common for all BChl's and an advantage when the compound is being used in PDT. The experimental absorption maxima of Tookad in toluene are positioned at 762, 535, 388, and 334 nm,⁹ as also indicated in Figure 3. For BPheid there are, to our knowledge, no experimental spectroscopic data for the higher energy excitations in toluene, and only the two lowest-energy absorption maxima, at 758 and 531 nm,⁴ are therefore indicated in Figure 4. From the spectra it can be concluded that ω B97XD displays the best positioned peak for the Q_y band for both Tookad and BPheid. The functional performance for the Tookad Q_y band is as follows: ω B97XD > CAM-B3LYP > M06-2X > M06 > M06HF > ω B97X > B3LYP > PBE1PBE > LC- ω PBE > ω B97. For BPheid the order is ω B97XD > M06HF > CAM-B3LYP > ω B97X > M06-2X > M06 > LC- ω PBE > ω B97 > B3LYP > PBE1PBE. For BPheid, CAM-B3LYP and M06HF were equally close to the experimental value ($\Delta\lambda = 19$ and 20 nm, respectively) but on opposite sides. The finding that CAM-B3LYP performs better than B3LYP on the lowest-lying excitation is supported by studies on Zn-bacteriochlorin and bacteriochlorin⁵⁵ as well as Zn-porphyrin in explicit aqueous solution,⁵⁶ in which the excitation energies are reduced when using CAM-B3LYP.

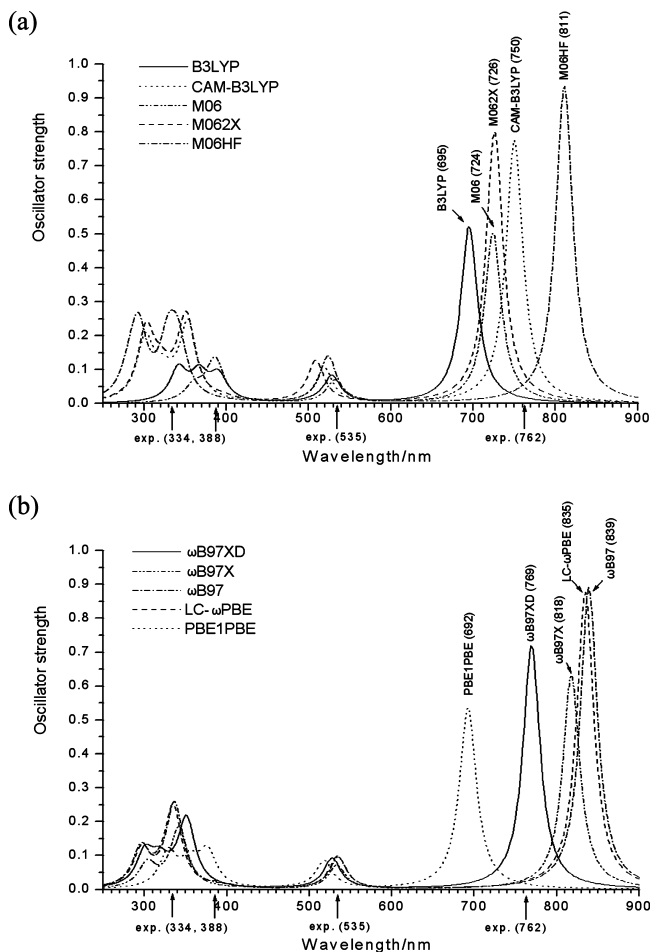


Figure 3. Absorption spectra of Tookad generated by (a) the B3LYP and M06 functional series and (b) the PBE and ω B97 functional series. Experimental values are displayed along the x axis.

LC- ω PBE and ω B97 generate absorptions at far too long wavelengths in both the case of Tookad and BPheid, meaning that the excitation energy is significantly underestimated. The opposite was observed for B3LYP and PBE1PBE that overestimated the excitation energy considerably (i.e., generating the absorption peak at far too short wavelength). B3LYP performs the best for the higher energy excitations of Tookad; however, this functional displays three peaks in the 300–400 nm region (also found when using THF as solvent), instead of two that are seen experimentally. The first and third peaks correspond very well to the experimental peaks found at 334 and 388 nm, whereas the second peak has no experimental match. The other functionals display two peaks each, albeit blue-shifted compared with B3LYP and experimental data. All functionals reproduce the small peak at 535 nm very well for Tookad, whereas for BPheid the functionals do not perform equally well in this case. Excitations with very small oscillator strengths, too small to be detected in the spectra, are generated by all functionals between the B and Q bands.

The effect of the basis sets (applied on all atoms but Pd) was investigated in the case of Tookad, and the resulting spectra obtained using the 6-31G(d,p), 6-31+G(d,p), and 6-311+G(2d,2p) basis sets are shown in Figure 5. The excited-state calculations were performed in toluene using

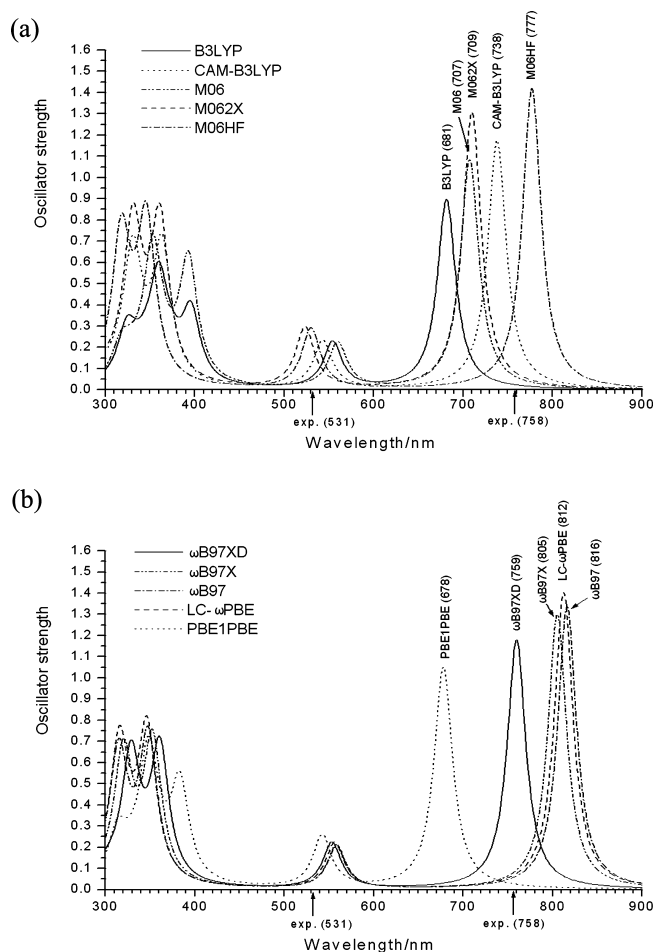


Figure 4. Absorption spectra of BPheid generated by (a) the B3LYP and M06 functional series and (b) the PBE and ω B97 functional series. Experimental values are displayed along the x axis.

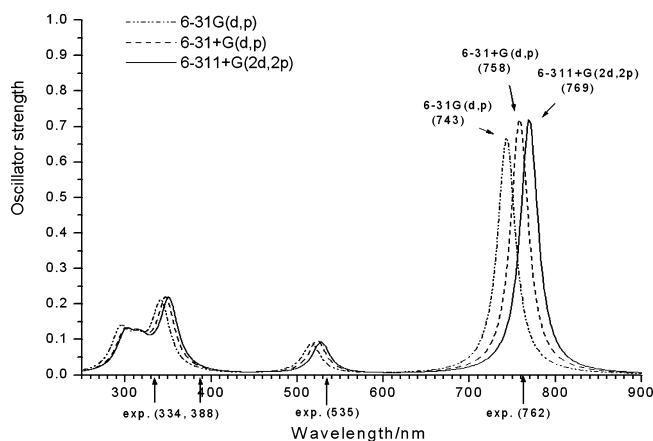
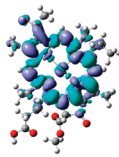
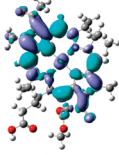
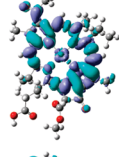
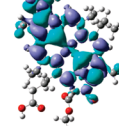


Figure 5. Absorption spectra for Tookad generated using ω B97XD in combination with three different basis sets. Experimental values are displayed along the x axis. Basis set for Pd is LanL2DZ throughout.

ω B97XD on the geometry generated at the B3LYP/6-31+G(d,p) level of theory. It can be seen that the larger basis sets with diffuse functions generate slightly better results, a difference that is seen mainly in the red-most region of the spectra. The basis set used in excited-state calculations has obviously a significantly smaller effect on the resulting

Table 3. Electron Density Differences between the Ground and Excited States of Tookad Generated Using the ω B97XD Functional

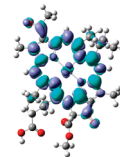
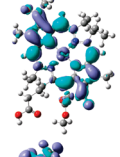
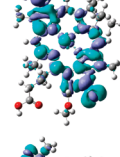
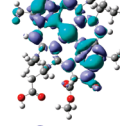
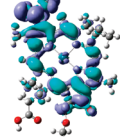
Excited state	Excitation energy/eV	Absorption/nm	Oscillator strength	Electron density difference between the ground state and excited state
1	1.6118	769	0.5396	
2	2.3503	528	0.1425	
9	3.5350	351	0.7192	
13	3.8974	318	0.1655	

spectra than the functional, a conclusion that has also been found previously in the case of Zn-porphyrin.^{29,30}

3.3. Electron Density Differences and Orbitals. As ω B97XD overall generated the most satisfying spectral results in this study, the computed electron density differences and orbitals of Tookad generated by this functional were investigated. For comparison the electron density differences and orbitals generated by B3LYP were also included. The 6-311+G(2d,2p) basis set was used together with toluene as the bulk solvent. The B3LYP geometry was used throughout. Electron density differences were calculated between the ground state and the excited states corresponding to the peaks displayed in the spectra. This means that four excited states were included for ω B97XD and five for B3LYP, as this functional generated an additional peak in the high-energy region of the spectra. The electron density differences for ω B97XD and B3LYP are displayed in Tables 3 and 4, respectively. Blue color represents a decrease in electron density and purple an increase in electron density of the excited state compared to the ground state. The density difference plots reveal that no dramatic restructuring of the electron distributions occur during the excitations; they are all essentially pure π - π^* excitations within the aromatic cores. Albeit a minor shift in electron density can be noted between the different metal d-orbitals, only the fourth excitation displayed involves any considerable interaction between the aromatic core and the metal.

The HOMO-4 to LUMO+4 orbitals are displayed in the Supporting Information. In agreement with the electron density difference plots, the shapes of the orbitals generated by the two functionals are almost identical. The only minor differences in the shapes are seen for LUMO+2 and LUMO+1. In LUMO+2 the d_{z^2} orbital of Pd in B3LYP is

Table 4. Electron Density Differences between the Ground and Excited States of Tookad Generated Using the B3LYP Functional

Excited state	Excitation energy/eV	Absorption/nm	Oscillator strength	Electron density difference between the ground state and excited state
1	1.7842	695	0.5220	
2	2.3473	528	0.1370	
8	3.1752	390	0.1397	
12	3.3815	367	0.1952	
14	3.6305	342	0.3930	

replaced by a $d_{x^2-y^2}$ orbital in ω B97XD. In LUMO+1 the conjugated π orbital of the ring system using B3LYP has a considerably smaller extension than when using the ω B97XD functional. The two lowest excitations (corresponding to the Q_y and Q_x bands) occur between HOMO and LUMO and between HOMO-1 and LUMO, respectively, for all functionals and for both Tookad and BPheid. However, as clearly seen in the absorption spectra, the excitation energies are obviously different. The energy gap between HOMO and LUMO is smaller when using ω B97XD compared with the same energy gap for B3LYP, which is also reflected in that ω B97XD overall generates the Q_y band at longer wavelengths than B3LYP. The higher energy excitations (the B bands) do not occur between the same orbitals for the different functionals, and those excitations involve more than only the two highest occupied and the two lowest unoccupied molecular orbitals, indicating that the four-orbital model does not correctly describe these transitions.

4. Conclusions

Tookad is a Pd-coordinated bacteriopheophorbide (Pd-BPheid) that has shown promising photodynamic properties, especially on prostate tumors. The low-lying excitations of Tookad and metal-free BPheid were studied computationally with the aim to reproduce the long-wavelength region of the absorption spectra through the use of time-dependent density functional theory (TD-DFT) in combination with new range-separated and meta hybrid density functional theory (DFT)

functionals. The commonly employed B3LYP functional was also included in the study, as it is well-known that this functional overestimates excitation energies.

B3LYP, M06HF, and ω B97XD were used for excited-state calculations on Tookad and BPheid geometries generated at the B3LYP and M06 level of theory, respectively, in toluene and tetrahydrofuran (THF). No significant difference was found in the calculated absorption spectra by the use of either B3LYP or M06 in geometry optimizations. For the excited-state calculations, ω B97XD was found to generate the Q_y transition band (the red-most absorption) closest to the experimental position, for both Tookad and BPheid. However, for the Q_x band (the second red-most absorption) the results are more inconsistent, with either B3LYP or ω B97XD generating the best results for Tookad and M06HF generating the best results for BPheid. In the case of Tookad, the three functionals do however perform highly equal for the Q_x band. When the CAM-B3LYP, M06, M06-2X, ω B97X, ω B97, LC- ω PBE, and PBE0 (PBE1PBE) functionals are included in the excited-state calculations of Tookad and BPheid in toluene, ω B97XD still performs the best for the Q_y band, followed by CAM-B3LYP. LC- ω PBE and ω B97 underestimate and B3LYP and PBE0 overestimate the Q_y band excitation energy considerably. These data hence differ from earlier benchmarking work, in which PBE0 was of consistently reasonable quality. A set of calculations with three different basis sets, with and without diffuse functions, indicated that the basis set has a minor effect on the calculated spectra. The overall accuracy for the Q_y band was 0–0.02 and 0–0.05 eV for Tookad and BPheid, respectively, for the ω B97XD functional.

We emphasize that the study is conducted on a limited system and that further work is needed in order to find the optimal combination of functionals and basis sets for optimizations and excitations. Clear from the current study is, however, that the evaluation of low-lying excitations is less straightforward to obtain at high accuracy than those in the UV region of the spectrum and that functionals that from a rational perspective cannot be justified for the current property seem to perform surprisingly well whereas others developed particularly with valence excitations in mind do not. It is also concluded that, whereas statistical treatment on a very large set of structurally divergent molecules may favor certain functionals, when focusing in on a particular class of compounds or a specific wavelength range, the results may be entirely different.

Acknowledgment. The Faculty of Science and Technology at Örebro University and the National University of Ireland - Galway are gratefully acknowledged for financial support.

Supporting Information Available: Cartesian coordinates for geometries of Tookad optimized in gas phase, THF and toluene using B3LYP and M06 functionals. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Henderson, B. W.; Sumlin, A. B.; Owczarczak, B. L.; Dougherty, T. J. *J. Photochem. Photobiol., B* **1991**, *10*, 303–313.
- Dougherty, T. J. Bacteriochlorophyll-A derivatives useful in photodynamic therapy. U.S. Patent 5,171,741, Dec 15, 1992.
- Hartwich, G.; Fiedor, L.; Simonin, I.; Cmiel, E.; Schafer, W.; Noy, D.; Scherz, A.; Scheer, H. *J. Am. Chem. Soc.* **1998**, *120*, 3675–3683.
- Limantara, L.; Sakamoto, S.; Koyama, Y.; Nagae, H. *Photochem. Photobiol.* **1997**, *65*, 330–337.
- Noy, D.; Fiedor, L.; Hartwich, G.; Scheer, H.; Scherz, A. *J. Am. Chem. Soc.* **1998**, *120*, 3684–3693.
- Geskes, C.; Hartwich, G.; Scheer, H.; Mantele, W.; Heinze, J. *J. Am. Chem. Soc.* **1995**, *117*, 7776–7783.
- Sundholm, D. *Chem. Phys. Lett.* **1999**, *302*, 480–484.
- Sundholm, D. *Chem. Phys. Lett.* **2000**, *317*, 545–552.
- Musewald, C.; Hartwich, G.; Pollinger-Dammer, F.; Lossau, H.; Scheer, H.; Michel-Beyerle, M. E. *J. Phys. Chem.* **1998**, *102*, 8336–8342.
- Vakrat-Haglilili, Y.; Weiner, L.; Brumfeld, V.; Brandis, A.; Salomon, Y.; McIlroy, B.; Wilson, B. C.; Pawlak, A.; Rozanowska, M.; Sarna, T.; Scherz, A. *J. Am. Chem. Soc.* **2005**, *127*, 6487–6497.
- Scherz, A.; Salomon, Y.; Brandis, A.; Scheer, H.; Palladium-substituted bacteriochlorophyll derivatives and use thereof. U.S. Patent 6,569,846, May 27, 2003.
- Chen, Q.; Huang, Z.; Luck, D.; Beckers, J.; Brun, P. H.; Wilson, B. C.; Scherz, A.; Salomon, Y.; Hetzel, F. W. *Photochem. Photobiol.* **2002**, *76*, 438–445.
- Schreiber, S.; Gross, S.; Brandis, A.; Harmelin, A.; Rosenbach-Belkin, V.; Scherz, A.; Salomon, Y. *Int. J. Cancer* **2002**, *99*, 279–285.
- Koudinova, N. V.; Pinthus, J. H.; Brandis, A.; Brenner, O.; Bendel, P.; Ramon, J.; Eshhar, Z.; Scherz, A.; Salomon, Y. *Int. J. Cancer* **2003**, *104*, 782–789.
- Borle, F.; Radu, A.; Fontollet, C.; van den Bergh, H.; Monnier, P.; Wagnieres, G. *Br. J. Cancer* **2003**, *89*, 2320–2326.
- Borle, F.; Radu, A.; Monnier, P.; van den Bergh, H.; Wagnieres, G. *Photochem. Photobiol.* **2003**, *78*, 377–383.
- Preise, D.; Mazor, O.; Koudinova, N.; Liscovitch, M.; Scherz, A.; Salomon, Y. *Neoplasia* **2003**, *5*, 475–480.
- Trachtenberg, J.; Bogaards, A.; Weersink, R. A.; Haider, M. A.; Evans, A.; McCluskey, S. A.; Scherz, A.; Gertner, M. R.; Yue, C.; Appu, S.; Aprikian, A.; Savard, J.; Wilson, B. C.; Elhilali, M. *J. Urol.* **2007**, *178*, 1974–1979.
- Haider, M. A.; Davidson, S. R. H.; Kale, A. V.; Weersink, R. A.; Evans, A. J.; Toi, A.; Gertner, M. R.; Bogaards, A.; Wilson, B. C.; Chin, J. L.; Elhilali, M.; Trachtenberg, J. *Radiology* **2007**, *244*, 196–204.
- Weersink, R. A.; Forbes, J.; Bisland, S.; Trachtenberg, J.; Elhilali, M.; Brun, P. H.; Wilson, B. C. *Photochem. Photobiol.* **2005**, *81*, 106–113.
- Trachtenberg, J.; Weersink, R. A.; Davidson, S. R. H.; Haider, M. A.; Bogaards, A.; Gertner, M. R.; Evans, A.; Scherz, A.; Savard, J.; Chin, J. L.; Wilson, B. C.; Elhilali, M. *BJU Int.* **2008**, *102*, 556–562.

- (22) Brun, P. H.; DeGroot, J. L.; Dickson, E. F. G.; Farahani, M.; Pottier, R. H. *Photochem. Photobiol. Sci.* **2004**, *3*, 1006–1010.
- (23) Tozer, D. J.; Amos, R. D.; Handy, N. C.; Roos, B. O.; Serrano-Andres, L. *Mol. Phys.* **1999**, *97*, 859–868.
- (24) Perpète, E. A.; Wathélet, V.; Preat, J.; Lambert, C.; Jacquemin, D. *J. Chem. Theory Comput.* **2006**, *2*, 434–440.
- (25) Andzelm, J.; Rinderspacher, B. C.; Rawlett, A.; Dougherty, J.; Baer, R.; Govind, N. *J. Chem. Theory Comput.* **2009**, *5*, 2835–2846.
- (26) Jacquemin, D.; Wathélet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435.
- (27) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2010**, *6*, 1532–1537.
- (28) Palma, M.; Cardenas-Jiron, G. I.; Rodriguez, M. I. M. *J. Phys. Chem. A* **2008**, *112*, 13574–13583.
- (29) Nguyen, K. A.; Day, P. N.; Pachter, R. *J. Chem. Phys.* **1999**, *110*, 9135–9144.
- (30) Nguyen, K. A.; Pachter, R. *J. Chem. Phys.* **2001**, *114*, 10757–10767.
- (31) Gouterman, M. *J. Mol. Spectrosc.* **1961**, *6*, 138–163.
- (32) Houssier, C.; Sauer, K. *J. Am. Chem. Soc.* **1970**, *92*, 779–791.
- (33) Fragata, M.; Norden, B.; Kurucsev, T. *Photochem. Photobiol.* **1988**, *47*, 133–143.
- (34) Hasegawa, J.; Ozeki, Y.; Ohkawa, K.; Hada, M.; Nakatsuji, H. *J. Phys. Chem. B* **1998**, *102*, 1320–1326.
- (35) Parusel, A. B. J.; Grimme, S. *J. Phys. Chem. B* **2000**, *104*, 5395–5398.
- (36) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (37) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (38) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.
- (39) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.
- (40) Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106.
- (41) Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P. *J. Chem. Phys.* **2007**, *126*, 154109.
- (42) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (43) Chai, J. D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084–106.
- (44) Chai, J. D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.
- (45) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *120*, 215–241.
- (46) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem.* **2006**, *110*, 5121–5129.
- (47) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem.* **2006**, *110*, 13126–13130.
- (48) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian09*, Revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.
- (49) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270–283.
- (50) Mennucci, B.; Cammi, R.; Tomasi, J. *J. Chem. Phys.* **1998**, *109*, 2798–2807.
- (51) Chipman, D. M. *J. Chem. Phys.* **2000**, *112*, 5558–5565.
- (52) Casida, M. E. In *Recent Advances in Density Functional Methods, Part 1*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; p 155–192.
- (53) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218–8224.
- (54) Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R. *J. Chem. Phys.* **1998**, *108*, 4439–4449.
- (55) Kobayashi, R.; Amos, R. D. *Chem. Phys. Lett.* **2006**, *420*, 106–109.
- (56) Govind, N.; Valiev, M.; Jensen, L.; Kowalski, K. *J. Phys. Chem. A* **2009**, *113*, 6041–6043.

JCTC

Journal of Chemical Theory and Computation

A Fourier Transform Method for Generation of Anharmonic Vibrational Molecular Spectra

Ivan Ivani,[‡] Vladimír Baumruk,[‡] and Petr Bouř^{*,†}

Institute of Organic Chemistry and Biochemistry, Academy of Sciences, 166 10 Prague, Czech Republic, and Charles University, Faculty of Mathematics and Physics, Institute of Physics, Ke Karlovu 5, 12116, Prague, Czech Republic

Received March 18, 2010

Abstract: Accurate computations of vibrational energies and vibrational spectra of molecules require inclusion of the anharmonic forces. In standard computational protocols, this leads to a large vibrational Hamiltonian matrix that needs to be diagonalized. Spectral intensities are calculated for individual transitions separately. In this work, an alternate direct generation of the spectral curves is proposed, based on a temporal propagation of a trial vibrational wave function followed by the Fourier transformation (FT). The method was applied to model water dimer and fenchone molecules. Arbitrary resolutions could be achieved by longer-time propagations, although a smaller integration time step (~ 0.02 fs) was needed for accurate peak frequencies than previously found for similar time-dependent applications within the harmonic approximation. Acceptably accurate relative vibrational spectra intensities were obtained when many random vectors used in the propagations were averaged. For a model fenchone Hamiltonian, simulated Raman and Raman optical activity (ROA) spectral shapes compared well with those obtained by the classical approach. The algorithm is amendable to parallelization. The lack of the lengthy and computer-memory-demanding diagonalization thus makes the FT procedure especially convenient for spectral simulations of larger molecules.

I. Introduction

Simulations of vibrational spectra are necessary to understand experimental data, and to obtain extensive information about molecular structures and force fields. Particularly for peptides, nucleic acids, and other biologically relevant systems, the vibrational spectroscopy provides a valuable means for the monitoring of specific structural and conformational features.¹ Historically, first spectral analyses were carried out by empirical correlations of IR or Raman band frequencies with the geometry.² Later theoretical approaches were based on simplified vibrational calculations, e.g., through parametrized force fields (FFs).³ Today, precise and fast quantum mechanical computations⁴ provide the most flexible way for theoretical spectral analyses. In particular, the density functional theory approximations can be applied for larger

molecules, including intensity simulations for experiments with unpolarized as well as, for example, circularly polarized radiation.⁵

The harmonic approximation based on the second derivatives of the nuclear potential⁶ is sufficient for many applications. Any molecule behaves like a system of independent harmonic oscillators at the harmonic limit. Typically, spectra of large biopolymers (nucleic acids, peptides) are simulated with this assumption because of the low resolution, limited spectral range, inhomogeneous band broadening caused by the solvent and molecular dynamics, and limited precision of available force fields.¹ For better accuracy or more advanced applications, anharmonic potential parts need to be included.^{7–10} Beyond the harmonic model, computation of molecular vibrational energies is no more a black box method, but advanced computational schemes are needed, including vibrational configuration interaction (VCI),^{11–13} vibrational self-consistent field (VSCF),^{9,14–16} many-body perturbation theory (PT),^{17,18} vibrational coupled clusters,¹⁹ etc.

* Corresponding author e-mail: bour@uochb.cas.cz.

[†] Academy of Sciences.

[‡] Charles University.

The VCI scheme, where the wave function is expressed as a linear combination of harmonic oscillator functions, is probably the most universal and most straightforward procedure. Unlike for the VSCF and PT approaches, fundamental and combination energy levels and spectral transitions can be obtained at the same time. Although VCI may become impractical for large systems,^{18,20} it represents an important benchmark as it is, in principle, equivalent to the exact Schrödinger solution. Unfortunately, similarly as for the electronic configuration interaction (CI),²¹ the dimension of the Hamiltonian required for a reasonable result quickly grows with the size of the molecule. Unlike for the electronic problem, however, where only few lowest-energy states are usually needed, a large portion of the vibrational energy levels covering the spectrum is required for vibrations.

Thus, a complete diagonalization of the vibrational Hamiltonian is typically needed to provide the transition energies, corresponding peak positions, and wave functions (eigenvectors) bearing spectral intensities. The classical in-memory iteration diagonalization routines are most convenient for small and medium dimensions ($N < \sim 10^4$).^{22,23} These direct algorithms occupy computer memory that is approximately proportional to N^2 and require times that scale as N^3 . Larger matrices can be more conveniently diagonalized, at least partially, by so-called power iteration methods, often referred to as (Jacobi-)Davidson algorithms, which perform the actual diagonalization in an intermediate (Krylov) vector space.^{24–28} The actual eigenspace can be built from the largest or from the smallest eigenvalue. The matrix does not need to be stored in memory, and the algorithm is simple, requiring essentially many matrix–vector multiplications only. When the matrix is sparse (which is often the case with the harmonic oscillator basis and a polynomial anharmonic potential), multiplications by the zeros can easily be avoided.

As each vector has to be orthonormalized against the previous ones, however, complete Davidson diagonalizations become difficult for larger matrices. It is also important to point out that for many applications detailed eigenvalue information is not needed. In particular for condensed phase spectroscopy, calculated line intensities are often convoluted with Gaussian or Lorentzian bands of finite widths, to simulate the inhomogeneous line broadening present in the experiment. Already for medium-sized molecules, observable peaks are usually composed from many unresolved vibrational transitions. Line spectrum simulations thus appear superfluous, whereas it is the spectral envelope that is desirable for comparison with the experiment to relate the structure and spectral response.

Therefore, the Fourier methods (Figure 1) may be a better option for unresolved spectral shapes. Within the harmonic limit, for example, it can be shown that classical molecular dynamic trajectories provided exact quantum results.²⁹ Propagation of a fictitious wave function in an arbitrary time was previously proposed to diagonalize giant Hessians and to generate corresponding vibrational spectra instead.³⁰ For large molecules, the Fourier transformation was much faster than the conventional diagonalizations. The spectral profiles were obtained by propagation and averaging of many trial vectors. However, the methods required the harmonic shape

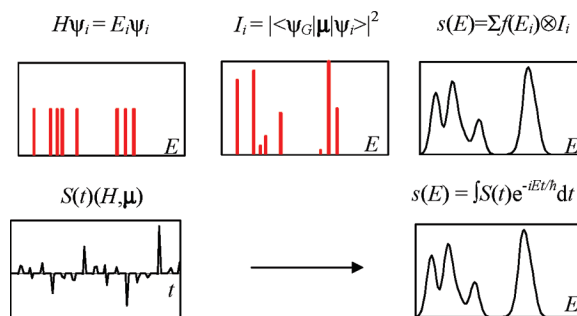


Figure 1. Schematic representation of the two processes of simulating vibrational spectra: (Top) By the usual way, discrete energies are found by a Hamiltonian (H) diagonalization; the intensities (I) are calculated from the eigenfunctions ψ and, for example, dipole moment μ , and the spectrum $s(E)$ is created by a convolution with an arbitrary peak shape f . (Bottom) Within the Fourier method, spectral function (S) develops in time, and the transformation provides the spectrum directly.

of the nuclear potential. In this work, we propose and test a different scheme suitable for a general anharmonic problem.

Time-dependent methods have always been popular in computational chemistry and were applied, for example, to simulations of the nuclear magnetic resonance,³¹ Raman scattering, infrared absorption, and vibrational circular dichroism.^{32,33} Anharmonic vibrational systems were also investigated; however, it should be noted that previous methods based on integration of classical trajectories do not provide all anharmonic corrections, such as the intermode coupling.^{32,34,35}

Modern mechanics—molecular mechanics (QM/MM) methods also facilitate computation of the spectra via time-dependent properties.^{36,37} In particular, more advanced spectroscopic experiments, such as the vibrational circular dichroism (VCD) or the two-dimensional (2D) spectroscopy, profit from various Fourier techniques.^{38–42} As a special class, the time-dependent filter-diagonalization methods²² make the spectral generation more efficient for a preselected frequency interval.⁴³ The methods are based on both classical^{44–46} and ab initio molecular dynamics trajectories^{47–49} but are mostly restricted to the harmonic potential.²⁹

Similarly, in the electronic spectroscopy and reactions, schemes like the multiconfigurational time-dependent Hartree approach⁵⁰ facilitate dynamic calculations for polyatomic molecules, a topic which goes beyond the scope of the present study. Rather than model real time-dependent processes, we introduce the time-dependent wave function and a spectral (e.g., dipole) function with the sole purpose of obtaining exact anharmonic energies and relative spectral intensities (including special polarized spectroscopies) for a general vibrational Hamiltonian. As the transition energies are needed rather than vibrational state energies, the exact ground state is obtained before the temporal propagation by the Davidson method. This, however, does not significantly increase the computational effort, unlike a complete Davidson diagonalization.

II. Theory

Consider a Hamiltonian H , wave functions $|K\rangle$, and energies E_K obli-ging the Schrödinger equation, $H|K\rangle = E_K|K\rangle$. A

general time-dependent wave function can be written as a sum, $\psi(t) = \sum_{K=1,N} d_K |K\rangle \exp(-iE_K t/\hbar)$, and propagated according to the time-dependent Schrödinger equation, $i\hbar\dot{\psi}(t) = H\psi(t)$, where \hbar is the Planck constant, d_K represents expansion coefficients, and N is the number of the basis functions. In the discrete time integration scheme detailed below, we calculated the wave function at time $t + dt$ as

$$\psi(t + dt) \cong \psi(t) + \dot{\psi}(t)dt + \frac{1}{2}\ddot{\psi}(t)dt^2 \quad (1)$$

with $\dot{\psi}(t) = H/(i\hbar)\psi(t)$, and $\ddot{\psi}(t)dt^2 = \psi(t) + \psi(t - 2dt) - 2\psi(t - dt)$; the wave function was renormalized at each time step.

The vibrational ground state $|G\rangle$ can easily be obtained by the Davidson diagonalization,^{25,51} as the first eigenvector. As pointed out in the Introduction, because the diagonalization becomes very inefficient for a large amount of required vectors,^{30,51} temporal propagations will be used to obtain spectral intensities coming from the remaining states instead.

The ground state wave function, besides the numerical propagation (eq 1), can also be propagated analytically as $\psi_G(t) = |G\rangle \exp(-iE_G t/\hbar)$, where E_G is the ground state energy. Additionally, we propagate a random function R and, for example, a dipole integral for the absorption spectrum

$$\boldsymbol{\mu}_R(t) = \langle R^*(t) | \hat{\boldsymbol{\mu}} | \psi_G(t) \rangle \quad (2)$$

where $\hat{\boldsymbol{\mu}}$ is the dipole moment operator. Adaptations for other spectral types are described below. The vector can always be thought of as decomposed to the exact solutions, $R(0) = \sum_K d_K^R \psi_K(0)$, where d_K^R represents unknown coefficients, so that

$$\boldsymbol{\mu}_R(t) = \sum_K d_K^{R*} \langle K | \hat{\boldsymbol{\mu}} | G \rangle e^{i\omega_{KG}t} \quad (3)$$

$\omega_{KG} = (E_K - E_G)/\hbar$, which can be Fourier-transformed to

$$\boldsymbol{\mu}_R(\omega) = \int \boldsymbol{\mu}_R(t) e^{-i\omega t} dt = 2\pi \sum_K d_K^{R*} \langle K | \hat{\boldsymbol{\mu}} | G \rangle \delta(\omega_{KG} - \omega) \quad (4)$$

Next, we define the absorption spectrum as

$$I_R(\omega) = \frac{\sqrt{2\pi}dN\omega}{4\pi^2} \left| \boldsymbol{\mu}_R(\omega) \right|^2 = \sum_K \langle K | \hat{\boldsymbol{\mu}} | G \rangle \cdot \langle G | \hat{\boldsymbol{\mu}} | K \rangle \omega \delta(\omega_{KG} - \omega) \quad (5)$$

In the derivation of eq 5 from 4, we used $\delta(\omega_{KG'} - \omega)\delta(\omega_{K'G} - \omega) \approx 1/(d\sqrt{2\pi})\delta_{KK'}\delta(\omega_{KG'} - \omega)$, which is valid for approximate "d functions" in a form of Gaussian bands, with a bandwidth d , $\delta_d(\omega) \approx \exp(-\omega^2/d^2)/(d\sqrt{\pi})$.

In order to remove the dependence on the choice of the initial vector R , the unknown state weights were replaced by the average, $|d_K^R|^2 \approx 1/N$. Note, that although the averaging was realized for expanding the vector to the harmonic oscillator basis, $R(0) = \sum_i r_i \varphi_i$, average expansion coefficients for any other orthogonal basis (in this case, the states ψ_K) are the same: Indeed, as the two $\{\varphi_i\}$ and $\{\psi_i\}$ sets are complete, we can always write $r_i = \sum_j d_j^R U_{ij}$, where \mathbf{U} is a

unitary transformation (rotation) matrix. For uncorrelated random numbers d_j^R within the interval $(-1,1)$, we obtain $\langle d_j^R d_l^R \rangle = \langle d_j^{R2} \rangle \delta_{jl}$, so that $\langle r_i^2 \rangle = \langle d_j^{R2} \rangle$. In other words, the averaging in any basis set provides the same final distribution.

Many random functions R_m ($m = 1-M$) were propagated to average the resultant intensities. Then, if the absorption index is defined as

$$\varepsilon(\omega) = (9.184 \times 10^{-3} M)^{-1} \sum_{R=1,M} I_R(\omega) \quad (6)$$

the dipole strength of each resolved transition $G \rightarrow K$ is equal to the usual relation⁵² $D_{KG} = 9.184 \times 10^{-3} \int \varepsilon d\omega/\omega$, where $D_{KG} = \langle K | \hat{\boldsymbol{\mu}} | G \rangle \cdot \langle G | \hat{\boldsymbol{\mu}} | K \rangle$ is in debye² and ε is in L mol⁻¹ cm⁻¹. In practical simulations, however, we used scaling of the calculated intensities by an empirical factor, based on a comparison of integrated IR and Raman intensities (calibrated for the water dimer). This procedure would eliminate the deviation of the simulated bands from ideal Gaussian functions. It should also be noted that exact absolute intensity simulations are not needed in most applications, as the relative band intensities bear most of the structural information.

The model vibrational Hamiltonian was chosen as

$$H = \frac{1}{2} \sum_{i=1}^{3n} (P_i^2 + \omega_i^2 Q_i^2) + \frac{1}{6} \sum_{i=1}^{3n} \sum_{j=1}^{3n} \sum_{k=1}^{3n} c_{ijk} Q_i Q_j Q_k + \frac{1}{24} \sum_{i=1}^{3n} \sum_{j=1}^{3n} \sum_{k=1}^{3n} \sum_{l=1}^{3n} d_{ijkl} Q_i Q_j Q_k Q_l \quad (7)$$

where $P_i = -i\hbar\partial/Q_i$, Q_i is normal mode coordinate, ω_i is the fundamental frequency, and n is the number of atoms. All cubic (c_{ijk}) and semidiagonal quartic (d_{ijkl} etc.; at least two indices were the same) constants were included. The size of the Hamiltonian was controlled by skipping the lowest-frequency modes and by considering harmonic states φ_i that significantly interact with the ground or fundamental (F) vibrations ($(|\langle \varphi_i | V | F \rangle| / (E_i - E_F)) \geq \text{threshold}$, where V represents the two last sums in eq 7). The threshold was set to 0 for the water dimer (all 0–5 × excited states included), and to 0.01 by default for the fenchone molecule. For the dimer, all modes were included, while for fenchone the six lowest modes were ignored. Only nonzero elements of \mathbf{H} were stored in memory.

III. Implementation

The algorithm derived above was implemented within the S4⁵³ Fortran code as follows:

(1) Calculate the Cartesian dipole derivatives $\boldsymbol{\mu}_R = \partial\boldsymbol{\mu}/\partial\mathbf{R}$; if required, calculate also the second dipole derivatives $\boldsymbol{\mu}_{RR} = \partial^2\boldsymbol{\mu}/(\partial\mathbf{R}\partial\mathbf{R})$, by a numerical differentiation. The Gaussian⁵⁴ program was used for the ab initio computations.

(2) Transform the first (second) derivatives into the normal mode coordinates, using the Cartesian-normal mode transformation ($3n \times 3n$) matrix \mathbf{S} , $\boldsymbol{\mu}_Q = \mathbf{S} \cdot \boldsymbol{\mu}_R$ ($\boldsymbol{\mu}_{QQ} = \mathbf{S}' \cdot \boldsymbol{\mu}_{RR} \cdot \mathbf{S}$).

(3) Construct the vibrational Hamiltonian matrix \mathbf{H} in the $N \times N$ harmonic oscillator basis $\{\varphi_i\}$, $i = 1-N$.

(4) Calculate the ground eigenvector \mathbf{g} ($|G\rangle = \sum_i g_i \varphi_i$) fulfilling $\mathbf{H} \cdot \mathbf{g} = E_G \mathbf{g}$, by the Davidson iteration.

(5) Precalculate the dipole matrix \mathbf{u} , $\mathbf{u}_i(0) = \sum_j g_j \langle \varphi_j | \hat{\boldsymbol{\mu}} | \varphi_i \rangle$,

where $\hat{\boldsymbol{\mu}} = \sum_{i=1}^{3n} \boldsymbol{\mu}_{Q_i} Q_i + 1/2 \sum_{i=1}^{3n} \sum_{j=1}^{3n} \boldsymbol{\mu}_{Q_i Q_j} Q_i Q_j$ is the vibrational dipole.

(6) Initialize the complex dipole function in the frequency domain (on a grid, typically 2000 points within 0–4000 cm^{-1}), $\boldsymbol{\mu}(\omega) = 0$, set time $t = 0$, and iteration step $k = 0$. In a set of complex random vectors \mathbf{r}_m ($m = 1-M$), set each component $r_{m,i}$ ($i = 1-N$) to a random number within $(-1$ to $+1)$ and normalize, so that $|\mathbf{r}_m| = 1$.

(7) Increment time t by dt and obtain:

$$\text{New vectors } \mathbf{r}_m^{(k+1)} = \mathbf{r}_m^{(k)} - (i/\hbar) \mathbf{H} \cdot \mathbf{r}_m^{(k)} + 1/2 \mathbf{d}2_m^{(k)}.$$

$$\text{Updated second derivatives } \mathbf{d}2_m^{(k+1)} = (\mathbf{r}_m^{(k)} + \mathbf{r}_m^{(k-2)} - 2\mathbf{r}_m^{(k-1)})/dt.$$

Dipoles $\boldsymbol{\mu}_m(t) = \mathbf{r}_m \cdot \mathbf{u} \exp(-iE_G t/\hbar)$. The scalar products in step 7 are related to the HO index, spanning $1-N$.

(8) Accumulate the dipole spectrum $\boldsymbol{\mu}(\omega) = \boldsymbol{\mu}(\omega) + e^{-i\omega t} \boldsymbol{\mu}_m(t) dt$, for each m .

(9) If $k < k_{\max}$, goto 7.

(10) From $\boldsymbol{\mu}(\omega)$, calculate the intensity according to eqs 5 and 6.

Modifications for Other Spectral Types. The algorithm above was derived for infrared absorption intensities. For vibrational circular dichroism (VCD), in steps 1 and 2, we additionally need to calculate Cartesian ($\mathbf{m}_C = \partial \mathbf{m}/\partial \mathbf{p}$, atomic axial tensor, AAT) and, consequently, normal mode ($\mathbf{m}_Q = \partial \mathbf{m}/\partial \mathbf{P}$) derivatives of the magnetic dipole moment \mathbf{m} ,⁵⁵ where \mathbf{p} and \mathbf{P} are the respective nuclear and normal mode momenta. The second-order anharmonic contribution was neglected for VCD and other spectral types. In step 5, besides matrix \mathbf{u} , we calculate $\mathbf{m}_i(\omega) = \sum_j g_j \langle \varphi_j | \mathbf{m} | \varphi_i \rangle$, where $\mathbf{m} = \mathbf{m}_Q \cdot \mathbf{P}$ is the vibrational magnetic dipole. The dipoles $\mathbf{m}_m(t) = \mathbf{r}_m \cdot \mathbf{m} \exp(-iE_G t/\hbar)$ are propagated in steps 6–9 for each random vector, and a frequency function $\mathbf{m}_m(\omega)$ is obtained in analogy to the electric dipole. The VCD spectrum corresponding to each m vector is $I_m(\omega) = [\sqrt{(2\pi) dN\omega}/(4\pi^2)] \text{Im}(\boldsymbol{\mu}_m^*(\omega) \cdot \mathbf{m}_m(\omega))$.

Raman spectra for various experimental setups can be obtained in a similar way, by replacing the dipole operator $\hat{\boldsymbol{\mu}} = \sum_{i=1}^{3n} \boldsymbol{\mu}_{Q_i} Q_i + 1/2 \sum_{i=1}^{3n} \sum_{j=1}^{3n} \boldsymbol{\mu}_{Q_i Q_j} Q_i Q_j$ by electric polarizability, $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_Q \cdot \mathbf{Q} + 1/2 \mathbf{Q} \cdot \boldsymbol{\alpha}_{QQ} \cdot \mathbf{Q}$. For backscattering Raman intensity,^{55,56} for example, we get $I_{R,180}(\omega) = K/(1 - \exp(-\omega/kT)) \sum_{\alpha=1-3} \sum_{\beta=1-3} \text{Re}(7\boldsymbol{\alpha}_{R,\alpha\beta}^*(\omega) \boldsymbol{\alpha}_{R,\alpha\beta}(\omega) + \boldsymbol{\alpha}_{R,\alpha\alpha}(\omega)^* \boldsymbol{\alpha}_{R,\beta\beta}(\omega))$. The constant K was chosen to be 1 (note that absolute intensities are rarely measured); k is the Boltzmann constant and T the temperature. The exponential factor accounts for scattering from excited vibrational levels as derived in the harmonic limit.⁵⁶ An alternative more exact path, based on individual low-energy states, used instead of the ground state and transitions weighed by the Boltzmann population, was not attempted. In that case, the temperature factor would have been omitted. However, anharmonic spectral correction in the lowest-wavenumber region, most affected by the temperature, is for most molecules rather small, and the harmonic-like temperature correction is thus sufficient.

By replacing the dipole operator by the electric dipole–magnetic dipole polarizability, $\hat{\mathbf{G}}' = \mathbf{G}'_Q \cdot \mathbf{Q} + 1/2 \mathbf{Q} \cdot \mathbf{G}'_{QQ} \cdot \mathbf{Q}$ (also referred to as the optical rotation tensor), and the electric dipole–electric quadrupole polarizability, $\hat{\mathbf{A}} = \mathbf{A}_Q \cdot \mathbf{Q} + 1/2 \mathbf{Q} \cdot \mathbf{A}_{QQ} \cdot \mathbf{Q}$, we can calculate Raman optical activity. The

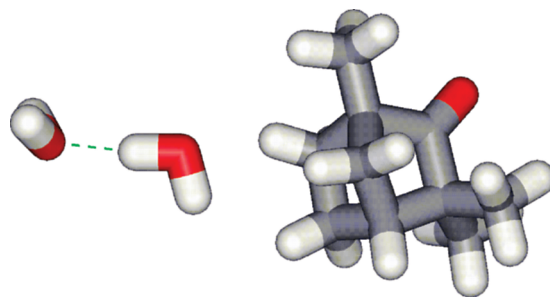


Figure 2. Water dimer and the fenchone molecule B3LYP/6-311++G** geometries.

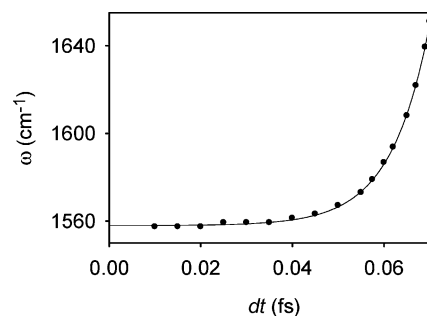


Figure 3. Dependence of the water HOH bending band frequency on the integration step, for water dimer.

backscattering incident circular polarized light intensity⁵⁵ was obtained as

$$\Delta I_{R,180}(\omega) = \frac{8K}{1 - \exp(-\omega/kT)} \sum_{\alpha=1-3} \sum_{\beta=1-3} \text{Re}(3\boldsymbol{\alpha}_{R,\alpha\beta}^*(\omega) \mathbf{G}'_{R,\alpha\beta}(\omega) - \boldsymbol{\alpha}_{R,\alpha\alpha}^*(\omega) \mathbf{G}'_{R,\beta\beta}(\omega) + \sum_{\delta=1-3} \sum_{\gamma=1-3} \varepsilon_{\alpha\gamma\delta} \boldsymbol{\alpha}_{R,\alpha\beta}(\omega)^* \mathbf{A}_{R,\gamma\delta\beta}(\omega))$$

The B3LYP⁵⁷/6311++G** method was used to compute the energy derivatives and the intensity tensors, as implemented in the Gaussian program.⁵⁴ Water dimer and the fenchone molecule (Figure 2) in equilibrium geometries were used for the modeling. Model VCI Hamiltonians with dimensions of 1325 (water) and 49 584 (fenchone) were used by default for most calculations; for fenchone, dimensions of 180, 509, 1456, 3560, 5689, and 119 817 were additionally used for the timing tests.

IV. Results

For exact Fourier transformation, the peak positions^{23,58} in the ω spectrum are constant. As was shown before already for the harmonic case,²⁹ in practical numerical integrations, larger time steps lead to overestimation of the peak frequencies. Indeed, as shown in Figure 3, where the water dimer bending vibration frequency is plotted as a function of the integration time step, larger steps (>0.06 fs) introduce errors of over 100 cm^{-1} . Only for steps below ~ 0.02 fs does the frequency stabilize. This is a relatively small fraction of the period of the corresponding harmonic motion, $T = 2\pi/\omega \approx 21$ fs. For harmonic wave function propagations, longer integration steps of ~ 0.1 fs could be used.²⁹ For some computations, however, steps as large as 2.4 fs were

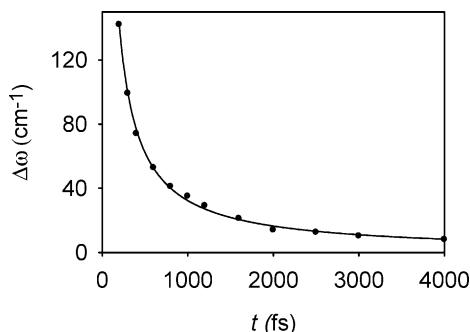


Figure 4. Dependence of the water dimer bending bandwidth on the integration time, for $dt = 0.02$ fs.

proposed.⁴³ We explain the need to use shorter integration steps for the anharmonic case even for lower-frequency states by a coupling to the higher-frequency states included in the Hamiltonian.

As follows from the general theory of Fourier transformation, the bandwidth is inversely proportional to the integration time, $\Delta\omega \sim t^{-1}$.^{23,58} This is also observed in the calculated dependence for the water dimer in Figure 4. As the width converges relatively slowly, the method does not seem to be usable for high-resolution spectra; in that case, many spectral points are additionally needed per frequency interval, which would further slow down the computations. On the other hand, the inhomogeneous band broadening is quite large for typical biomolecular spectra, on the order of ~ 20

cm^{-1} ,^{59,60} so that the propagation times can be limited. That means that for a 0.02 fs time step (used to achieve a high precision of central frequencies, cf. Figure 3), about $4000/0.2 = 200\,000$ propagation points are needed.

Although the spectral intensities that can be obtained with the FT method are only approximate, for a large number of the random vectors, relative band ratios are reasonably close to the exact result. This is documented in Figure 5, where backscattering Raman and ROA spectra of fenchone are simulated for M (number of the vectors) = 5, 10, and 50 and compared to exact intensities calculated by the direct diagonalization of the model $49\,584 \times 49\,584$ VCI Hamiltonian. Already for $M = 5$, the raw Raman spectral profile is similar to the direct calculation; the relative peak ratios are further improved for $M = 50$. The ROA signal converges more slowly, especially within the $1400\text{--}1600\text{ cm}^{-1}$ region, where many overlapped transitions (mostly C–H bending vibrations) are present. However, the simulation $M = 50$ provides the correct relative intensity and sign pattern for ROA, too. Both the Raman and ROA CH stretching higher-frequency signal seems to converge faster than that for vibrations below 2000 cm^{-1} . The calculated vibrational frequencies correspond reasonably well to the observed values;⁶¹ however, we leave a detailed comparison to the experimental spectral profiles for a future study because of the complexity of the problem.

As a more exact means to document the convergence, in Figure 6, part A, we plot an example of an actual r_{mi}

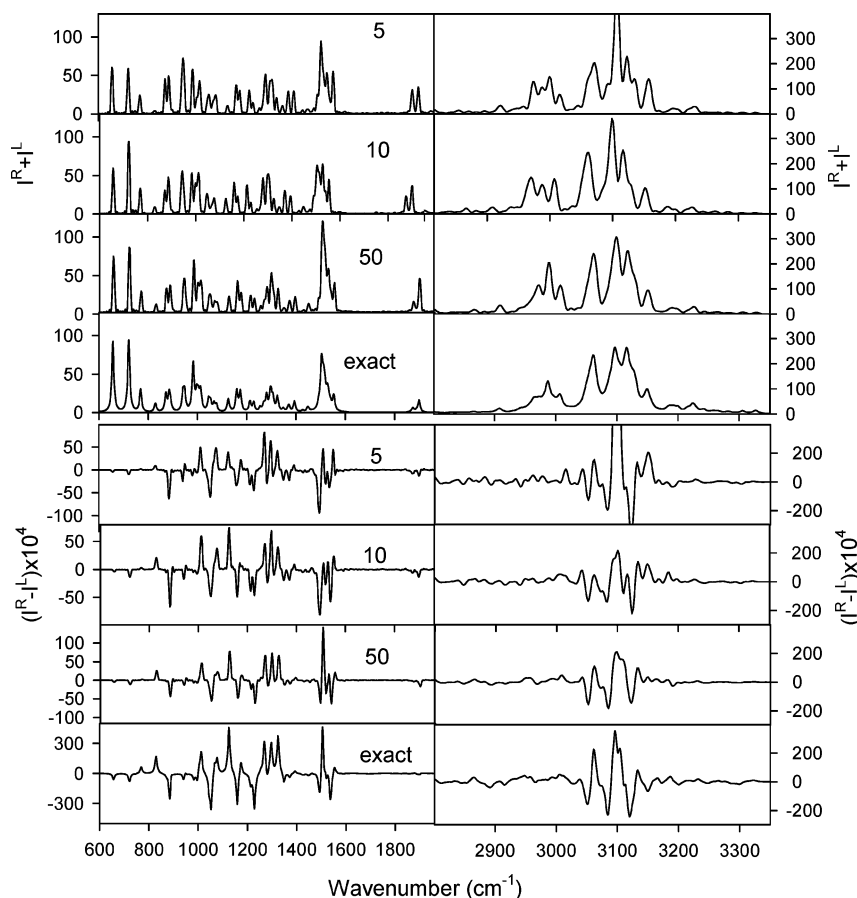


Figure 5. Dependence of the Raman (top) and Raman optical activity (bottom) spectra of fenchone on the number of random vectors used in the propagation.

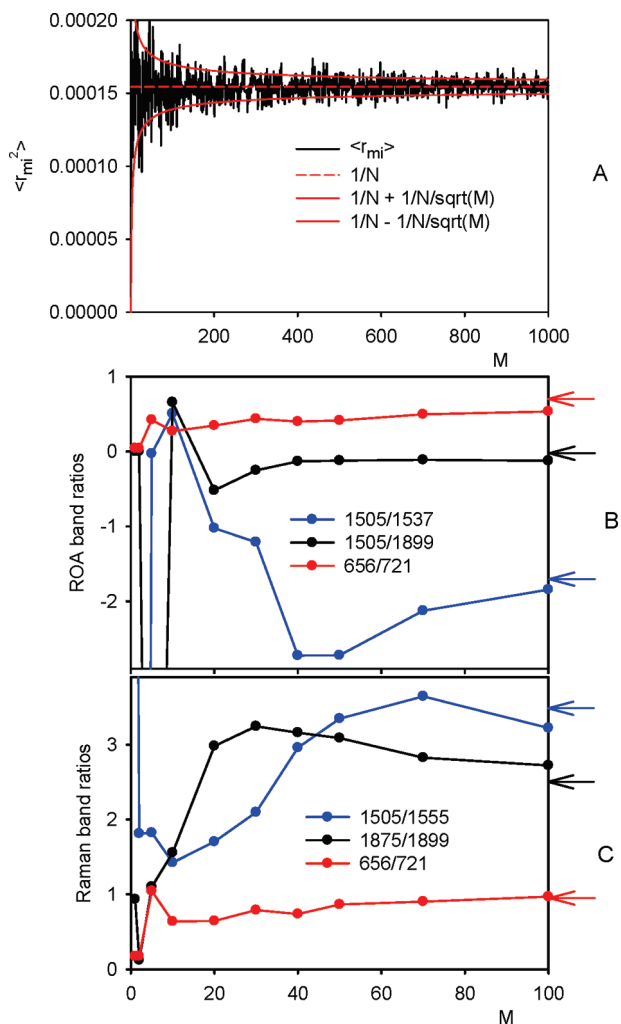


Figure 6. Convergence properties of spectral intensities on the number of random vectors: (A) average, RMS deviation interval⁶² ($N^{-1} \pm N^{-1}M^{-1/2}$), and actual values for a random coefficient ($i = 10$) for the fenchone simulation in Figure 5 with $N = 6475$ and ratios of selected (B) ROA and (C) Raman peak intensities. Central peak frequencies are indicated in cm^{-1} ; the arrows mark exact values.

coefficient averaging and the root-mean-square deviation that converges as $\sim 1/\sqrt{M}$.⁶² Although, as discussed above, we cannot get the actual state probabilities (d_i), from eqs 3–5, it is clear that the intensity will converge in the same manner that this factor does. The possible error proportional to the square root of M converges rather slowly; thus benchmark simulations with large values of M are clearly inefficient. On the other hand, in accord with the observation of the spectral convergence in Figure 5, a reasonable intensity error of $\sim 10\%$ can be obtained with a limited amount (<100) of the vectors, which is sufficient in many applications of the vibrational spectroscopy.

Actual convergence of the Raman and ROA band ratios (Figure 6, parts B and C) is more complicated due to the band overlaps; however, the trends are clearly given by the basic $1/\sqrt{M}$ dependence for the coefficients. From the figure, we also see that simulations with $M < 20$ should be avoided for ROA, as they may even lead to the wrong signs for some peaks. For the selected examples of three peak pairs in Figure

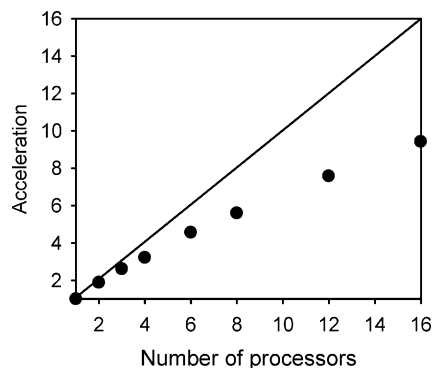


Figure 7. Dependence of the acceleration on the number of processors (fenchone IR spectrum calculation, pgf77-OMP-linux software environment, 4 Intel E7330/2.40 GHz CPUs on Supermicro X7QCE motherboard).

6, the lowest-energy lone-standing transitions ($656/721 \text{ cm}^{-1}$) converge most smoothly.

The number of vibrational degrees of freedom associated with the number of atoms does not seem to be important for the convergence properties; the water dimer spectra (not shown) behaved similarly to that of fenchone. However, as the density of vibrational states increases and the peaks became more overlapped in more complex molecules, higher accuracy, and thus presumably a larger number of the starting vectors, will be required for simulations on larger systems.

As observed also for other time-dependent approaches,^{29,43} it is difficult to extract information about the individual normal mode contribution to the spectrum. For harmonic potential, this is partially solvable by a specially designed propagation scheme.⁶³ In anharmonic computations, the concept of normal modes vanishes completely. However, in a majority of practical computations, the harmonic approximation is realistic enough to provide reliable information about the origin of observable transitions.

As the vectors can be propagated independently, the algorithm is amendable to parallelization. Our OMP shared memory implementation (<http://openmp.org>) did not lead to a perfect scaling (cf. Figure 7); nevertheless, it documents the significant speedups that can easily be achieved on common shared-memory multiprocessor computers. More importantly, the FT algorithm becomes very convenient for larger Hamiltonian dimensions. This is documented in Figure 8, where the diagonalization times needed for the direct and Davidson computations are compared to the FT simulations for variously sized fenchone VCI Hamiltonians. The Davidson method is apparently quite inefficient, and the CPU time rises steeply. The direct diagonalization is very fast for smaller matrices, but the N^3 time and N^2 memory scaling make it inconvenient for larger ones; for $N \sim 6000$, the FT methodology becomes the fastest scheme for the vibrational spectra generation. As pointed out above, slightly longer times are required for more resolved spectra (longer propagation needed) and more accurate spectral intensities (requiring many random vector averaging). Still, the FT method would be the most convenient when the Hamiltonian reaches a certain limit. Additionally, only nonzero Hamiltonian elements need to be stored for FT, unlike for the direct methods.

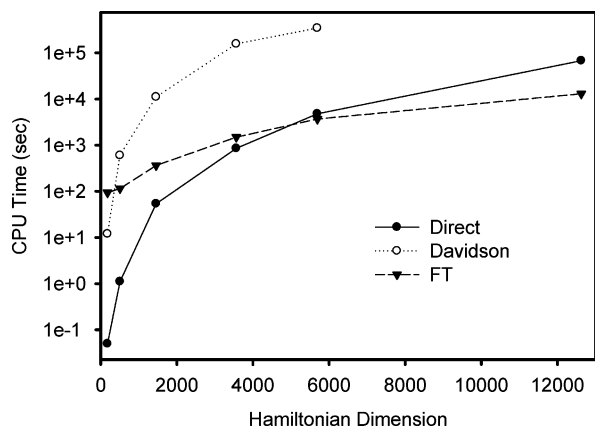


Figure 8. CPU time needed for the fenchone Hamiltonian diagonalization with the direct and Davidson methods, and the FT spectral generation, as a function of the matrix dimension, for one Intel E5530 2.4 GHz processor. The Davidson iteration was limited to wavenumbers below 2000 cm^{-1} ; the FT computation was done with 10 000 steps and one random vector only.

V. Conclusions

The proposed computational scheme enabled us to estimate conveniently vibrational spectral profiles based on the VCI Hamiltonian and intensity tensor derivatives. Because the ground state could be calculated by the classical Davidson method, the Fourier transformation with suitably chosen integration steps provided exact transition frequencies. Besides the wave function, electromagnetic tensors (e.g., the electric dipole for infrared absorption) were propagated, which enabled a simultaneous computation of spectral intensities. Only approximate absolute intensities could be simulated; however, propagation of many random vectors and the averaging led to faithful relative band intensities and correct ROA sign patterns, with accuracy sufficient for most molecular structural studies based on the vibrational spectra. For large molecules (large VCI Hamiltonians), the algorithm provided the spectra faster than the classical methods based on the explicit matrix diagonalization.

Acknowledgment. This work was supported by the Grant Agency of the Czech Republic (202/07/0732) and the Grant Agency of the Academy of Sciences (A400550702, M200550902).

References

- (1) Kubelka, J.; Bouř, P.; Keiderling, T. A. Quantum Mechanical Calculations of Peptide Vibrational Force Fields and Spectral Intensities. In *Advances in Biomedical Spectroscopy, Biological and Biomedical Infrared Spectroscopy*; Barth, A., Haris, P. I., Eds.; IOS Press: Amsterdam, 2009; Vol. 2, pp 178.
- (2) Califano, S. *Vibrational states*; John Wiley & Sons: London, 1976.
- (3) Narayanan, U.; Keiderling, T. A. *J. Am. Chem. Soc.* **1983**, *105*, 6406.
- (4) Pulay, P. Analytical derivative techniques and the calculation of vibrational spectra. In *Modern electronic structure theory*; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; Vol. 2, pp 1191.
- (5) Cheeseman, J. R.; Frisch, M. J.; Devlin, F. J.; Stephens, P. J. *Chem. Phys. Lett.* **1996**, *252*, 211.
- (6) Papoušek, D.; Aliev, M. R. *Molecular Vibrational/Rotational Spectra*; Academia: Prague, 1982.
- (7) Bounouar, M.; Scheurer, C. *Chem. Phys.* **2006**, *323*, 87.
- (8) Daněček, P.; Kapitán, J.; Baumruk, V.; Bednářová, L.; Kopecký, V., Jr.; Bouř, P. *J. Chem. Phys.* **2007**, *126*, 224513.
- (9) Hansen, M. B.; Sparta, M.; Seidler, P.; Toffoli, D.; Christiansen, O. *J. Chem. Theory Comput.* **2010**, *6*, 235.
- (10) Andrushchenko, V.; Matějka, P.; Anderson, D. T.; Kaminský, J.; Horníček, J.; Paulson, L. O.; Bouř, P. *J. Phys. Chem. A* **2009**, *113*, 9727.
- (11) Fujisaki, H.; Yagi, K.; Hirao, K.; Straub, J. E. *Chem. Phys. Lett.* **2007**, *443*, 6.
- (12) Chakraborty, A.; Truhlar, D. G.; Bowman, J. M.; Carter, S. *J. Chem. Phys.* **2004**, *121*, 2071.
- (13) Neff, M.; Rauhut, G. *J. Chem. Phys.* **2009**, *131*, 124129.
- (14) Bowman, J. M. *J. Chem. Phys.* **1978**, *68*, 608.
- (15) Gerber, R. B.; Ratner, M. A. *Chem. Phys. Lett.* **1979**, *68*, 195.
- (16) Bounouar, M.; Scheurer, C. *Chem. Phys.* **2008**, *347*, 194.
- (17) Norris, L. S.; Ratner, M. A.; Roitberg, A. E.; Gerber, R. B. *J. Chem. Phys.* **1996**, *105*, 11261.
- (18) Daněček, P.; Bouř, P. *J. Comput. Chem.* **2007**, *28*, 1617.
- (19) Christiansen, O. *J. Chem. Phys.* **2004**, *120*, 2149.
- (20) Christiansen, O.; Luis, J. M. *Int. J. Quantum Chem.* **2005**, *104*, 667.
- (21) Schleyer, P. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F., III; Schreiner, P. R. *The Encyclopedia of Computational Chemistry*; John Wiley & Sons: Chichester, U.K., 1998.
- (22) Grotendorst, J. *Modern methods and algorithms of quantum chemistry*; John von Neumann Institute for Computing: Jülich, Germany, 2000; Vol. 1.
- (23) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in Fortran*, 2nd ed.; Cambridge University Press: New York, 1992.
- (24) Kaledin, A. L.; Kaledin, M.; Bowman, J. M. *J. Chem. Theory Comput.* **2006**, *2*, 166.
- (25) Davidson, E. R. *J. Comput. Phys.* **1975**, *17*, 87.
- (26) Martins, J. L.; Cohen, M. L. *Phys. Rev. B* **1988**, *37*, 6134.
- (27) Mitin, A. V. *J. Comput. Chem.* **1994**, *15*, 747.
- (28) vanderVorst, H. A. *Iterative Krylov Methods for Large Linear Systems*; Cambridge University Press: Cambridge, U. K., 2003.
- (29) Horníček, J.; Kaprálová, P.; Bouř, P. *J. Chem. Phys.* **2007**, *127*, 084502.
- (30) Kubelka, J.; Bouř, P. *J. Chem. Theory Comput.* **2009**, *5*, 200.
- (31) Gordon, R. G. *J. Chem. Phys.* **1965**, *42*, 3658.
- (32) Noid, D. W.; Koszykowski, M. L.; Marcus, R. A. *J. Chem. Phys.* **1977**, *67*, 404.
- (33) Abbate, S.; Longhi, G.; Kwon, K.; Moscovitz, A. *J. Chem. Phys.* **1998**, *108*, 50.
- (34) Kim, H.; Rossky, P. J. *J. Chem. Phys.* **2006**, *125*, 074107.

- (35) Kinnaman, C. S.; Cremeens, M. E.; Romesberg, F. E.; Corcelli, S. A. *J. Am. Chem. Soc.* **2006**, *128*, 13334.
- (36) Hahn, S.; Lee, H.; Cho, M. *J. Chem. Phys.* **2004**, *121*, 1849.
- (37) Mankoo, P. K.; Keyes, T. *J. Chem. Phys.* **2006**, *124*, 204503.
- (38) Lee, K. K.; Hahn, S.; Oh, K. I.; Choi, J. S.; Joo, C.; Lee, H.; Han, H.; Cho, M. *J. Phys. Chem. B* **2006**, *110*, 18834.
- (39) Torii, H. *J. Phys. Chem. A* **2006**, *110*, 9469.
- (40) Gnanakaran, S.; Hochstrasser, R. M. *J. Am. Chem. Soc.* **2001**, *123*, 12886.
- (41) Loparo, J. J.; Roberts, S. T.; Tokmakoff, A. *J. Chem. Phys.* **2006**, *125*, 194521.
- (42) Krummel, A. T.; Zanni, M. T. *J. Phys. Chem. B* **2006**, *110*, 24720.
- (43) Silva, A. J. R.; Pang, J. W.; Carter, E. A.; Neuhauser, D. *J. Phys. Chem. A* **1997**, *102*, 881.
- (44) Liang, Y.; Miranda, C. R.; Scandolo, S. *J. Chem. Phys.* **2006**, *125*, 194524.
- (45) Yang, S.; Cho, M. *J. Phys. Chem. B* **2007**, *111*, 605.
- (46) Gorbunov, R. D.; Nguyen, P. H.; Kobus, M.; Stock, G. *J. Chem. Phys.* **2007**, *126*, 054509.
- (47) Yamauchi, Y.; Nakai, H. *J. Chem. Phys.* **2004**, *121*, 11098.
- (48) Putrino, A.; Parrinello, M. *Phys. Rev. Lett.* **2002**, *88*, 176401.
- (49) Seibt, J.; Engel, V. *J. Chem. Phys.* **2007**, *126*, 074110.
- (50) Meyer, H. D.; Le Quere, F.; Leonard, C.; Gatti, F. *Chem. Phys.* **2006**, *329*, 179.
- (51) Murray, C. W.; Racine, S. C.; Davidson, E. R. *J. Comput. Chem.* **1992**, *103*, 382.
- (52) Charney, E. *The Molecular Basis of Optical Activity*; Wiley-Interscience: New York, 1979.
- (53) Bouř, P. *S4*; Academy of Sciences: Prague, 1994–2010.
- (54) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
- (55) Barron, L. D. *Molecular Light Scattering and Optical Activity*; Cambridge University Press: Cambridge, U. K., 2004.
- (56) Polavarapu, P. L. *Vibrational Spectra and Structure* **1984**, *13*, 103.
- (57) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (58) Johnson, K. *Numerical Methods in Chemistry*; Marcel Dekker, Inc.: New York, 1980; Vol. 7.
- (59) Keiderling, T. A.; Kubelka, J.; Hilario, J. Vibrational circular dichroism of biopolymers. Summary of methods and applications. In *Vibrational spectroscopy of polymers and biological systems*; Braiman, M., Gregoriou, V., Eds.; CRC Press: Boca Raton, FL, 2006; pp 253.
- (60) Krimm, S. Vibrational spectroscopy of polypeptides. In *Modern Polymer Spectroscopy*; Zerbi, G., Ed.; Wiley-VCH: New York, 1999; pp 239.
- (61) Longhi, G.; Abbate, S.; Gangemi, R.; Giorgio, E.; Rosini, C. *J. Phys. Chem. A* **2006**, *110*, 4958.
- (62) Everitt, B. S. *The Cambridge Dictionary of Statistics*; Cambridge University Press: Cambridge, U. K., 2002; p 410.
- (63) Martines, M.; Gaigeot, M. P.; Borgis, D.; Vuilleumier, R. *J. Chem. Phys.* **2006**, *125*, 144106.

CT100150F

Singlet–Triplet States Interaction Regions in DNA/RNA Nucleobase Hypersurfaces

Remedios González-Luque, Teresa Climent, Israel González-Ramírez,
Manuela Merchán, and Luis Serrano-Andrés*

*Instituto de Ciencia Molecular, Universitat de València, Apartado 22085,
ES-46071 Valencia, Spain*

Received March 26, 2010

Abstract: The present study provides new insight into the intrinsic mechanisms for the population of the triplet manifold in DNA nucleobases by determining, at the multiconfigurational CASSCF/CASPT2 level, the singlet–triplet states crossing regions and the main decay paths for their lowest singlet and triplet states after near-UV irradiation. The studied singlet–triplet interacting regions are accessible along the minimum energy path of the initially populated singlet bright $^1\pi\pi^*$ state. In particular, all five natural DNA/RNA nucleobases have, at the end of the main minimum energy path and near a conical intersection of the ground and $^1\pi\pi^*$ states, a low-energy, easily accessible, singlet–triplet crossing region directly connecting the lowest singlet and triplet $\pi\pi^*$ excited states. Adenine, thymine, and uracil display additional higher-energy crossing regions related to the presence of low-lying singlet and a triplet $n\pi^*$ state. These funnels are absent in guanine and cytosine, which have the bright $^1\pi\pi^*$ state lower in energy and less accessible $n\pi^*$ states. Knowledge of the location and accessibility of these regions, in which the singlet–triplet interaction is related to large spin–orbit coupling elements, may help to understand experimental evidence such as the wavelength dependence measured for the triplet formation quantum yield in nucleobases and the prevalence of adenine and thymine components in the phosphorescence spectra of DNA.

1. Introduction

Phosphorescence spectra of DNA at low temperatures have been established as consisting of two basic components which originate mainly from thymine and, to a lesser extent, from adenine.^{1–3} Although triplet state formation and phosphorescence data of individual nucleobases and different derivatives in several media and conditions have been reported and reviewed,^{4–7} including recent studies employing external photosensitizers,^{8–10} the specifics of the intrinsic population mechanism of the triplet manifold in each of the nucleobases has not been understood so far. The different fates of their triplet states, explaining, for instance, the prevalence of two of the bases in the phosphorescence spectra of DNA, the absence of triplet guanine signals, or the triplet state involvement in the fast relaxation processes of nucleobases,

in particular for thymine,¹¹ have still to be elucidated. Triplet states of molecular systems are frequent intermediates in important photoinduced reactions. Both their usual biradical character and relatively long lifetimes make them reactive species prone to interacting with other compounds.¹² Triplet states of DNA/RNA purine and pyrimidine nucleobases are not an exception, and they have been determined to participate in UV-promoted photoreactions as the formation of phototherapeutic nucleobase-pharmakon adducts¹³ or the photodimerization of pyrimidine nucleobases, considered to be the most frequent genetic lesion taking place after UV-light irradiation.^{7,14–16} Since most of the recent attention has been focused on the rapid dynamics of the initially populated singlet states of DNA/RNA nucleobases,^{17–22} their intersystem crossing (ISC) mechanisms and triplet states' decay processes are only now starting to be analyzed.^{23–26} The present study aims to present a unified scheme, based on quantum chemical grounds, for the description of the main

* Corresponding author fax: (+34) 96-3544427, e-mail: Luis.Serrano@uv.es.

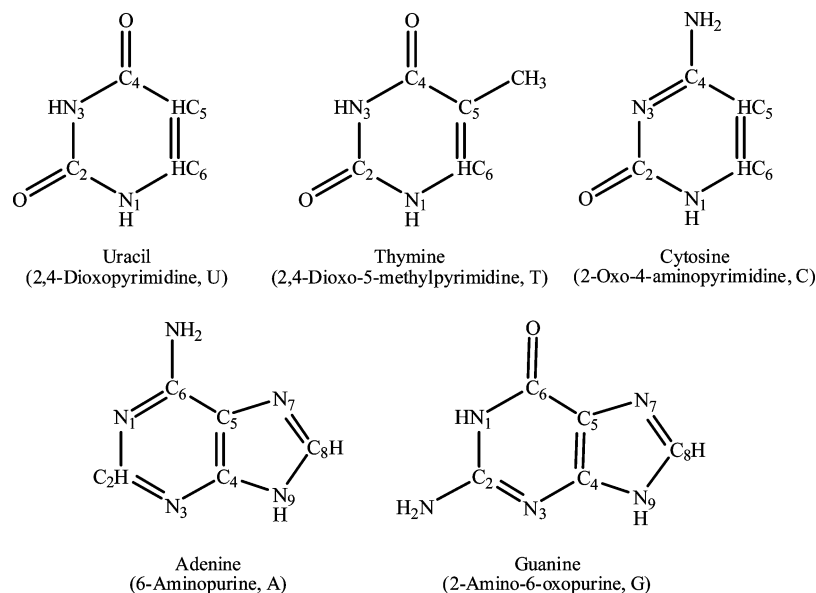


Figure 1. Structure, labeling, common name, IUPAC name, and acronym used for the five natural DNA/RNA nucleobases.

decay pathways for the singlet and triplet states of the five natural DNA/RNA nucleobases, thymine (T), uracil (U), cytosine (C), adenine (A), and guanine (G) (see Figure 1), locating the singlet–triplet crossing regions and computing the related spin–orbit coupling terms in order to provide insight into the intrinsic mechanisms of triplet state population in these molecules and to help rationalize the observed experimental data.

The triplet state population may proceed via endogenous or exogenous photosensitization from other triplet species or by efficient intersystem crossing (ISC) from the initially excited singlet state. There is an essential consensus that efficient radiationless transitions among states of the same multiplicity leading to internal conversion (IC) take place in the close vicinity of conical intersection (CI) regions and that the probability for the decay and the IC rates relate to the size of the nonadiabatic coupling matrix elements between the states.^{27–29} The situation is more complex for the computation of ISC rates. In this case, the efficiency of the interaction between states of different multiplicities, for instance, singlet–triplet, seems to be reasonably well described by the Fermi Golden Rule, which relates the strength of the interaction to the extent of the vibronic spin–orbit coupling (SOC) factors and the Franck–Condon (FC) weighted density of states.²⁷ Recent studies of Marian and co-workers^{26,30,31} have proved that the efficiency of an ISC process relies on a subtle balance of effects, including an enlarged density of vibrational states and a proper overlap of vibrational wave functions which, in turn, enhance the vibronic SOC effects. The decrease of the energy gap between singlet and triplet states, and in particular the presence of singlet–triplet degeneracies, crossing regions, especially when related to the existence of low-energy out of plane vibrational modes, is a good indication of a high density of states, and it is therefore conceivable that singlet–triplet crossings play an important role for increased ISC population transfer rates.¹² This relevance is well-known in the emerging field of multistate reactivity,^{27,32,33} in which the presence of singlet–triplet crossings and the occurrence

of corresponding ISC processes in the vicinity of the ground-state transition state regions become crucial for the enhancement of the reaction rates.³⁴ As they compete with generally faster internal conversion processes, intersystem crossings or spin crossovers can also be expected to be more efficient in energy trapping regions, for instance, near singlet states minima or sloped singlet–singlet conical intersections.^{35,36} Full reaction dynamics calculations including in a balanced and accurate way nonadiabatic and spin–orbit coupling effects for polyatomic systems like those considered here have not been performed yet. Until those studies are available, calculations of ISC rates in which the vibronic spin–orbit and overlap coupling effects are considered give the best information about the efficiency of the ISC process.^{26,30} Our goal in the present research is to determine the presence and accessibility of the singlet–triplet degeneracy regions in natural nucleobases along the main singlet decay pathways and provide hints of their relevance for ISC by computing also electronic SOC terms.

The strategy employed here starts by obtaining the minimum energy paths (MEPs) leading from the primary step of the photochemical process after UV light absorption in DNA nucleobases, being basically the population of the spectroscopically bright singlet excited state, here always the so-called $^1(\pi\pi^* L_a)$ state, toward the singlet–triplet degeneracy regions and finally the lowest-energy and reactive triplet excited state $^3(\pi\pi^* L_a)$, and calculating electronic SOC terms between relevant states.

Recent quantum-chemical *ab initio* CASPT2 studies have provided a unified model for the rapid internal conversion (IC) of the singlet excited DNA/RNA nucleobases manifold^{18,20,29,37–46} that allowed a proper rationalization of the experimental findings.^{17,47} The observed ultrafast decay component in all natural nucleobases, both in the gas phase and in solution, can be interpreted in terms of the barrierless character of the minimum energy path (MEP) associated with the lowest singlet state of the $\pi\pi^*$ type, $^1(\pi\pi^* L_a)$, toward a conical intersection (CI) with the ground state, (gs/ $\pi\pi^*$)_{CI}. Secondary decay paths involving the lowest $^1n\pi^*$ state and

even a higher $^1\pi\pi^*$ state have been also identified.^{20,23,29,38–41} Within the context of the photochemical reaction path approach⁴⁸ and the current theoretical paradigm for nonadiabatic photochemistry,^{28,29} it is possible to analyze how the lowest triplet state can be reached efficiently by finding the singlet–triplet crossing (STC) regions more easily accessed from the FC MEP on the $^1(\pi\pi^* L_a)$ state, which represents the major deactivation path responsible of the rapid IC process detected in the molecule. Further studies combining the calculation of ISC rates and wave packet evolution will have to determine how efficient actually are our proposed channels. The obtained results suggest that enhancements in the population yield of the lowest triplet state of the natural DNA/RNA nucleobases can be related to the presence in three of them, T, U, and A, of more ISC channels along the singlet state MEP, in particular those related with low-lying singlet and triplet $n\pi^*$ states that act as intermediate population switchers, unlike what occurs in C and G. The obtained scheme may help to understand how the intrinsic population of the lowest triplet state can take place in vacuo for all the nucleobases, why T and A triplet states seem to prevail on the DNA phosphorescence spectrum and can be expected to have a larger quantum yield of formation (ϕ_{ISC}) than the other nucleobases, and what the molecular basis is for the detected wavelength dependence of ϕ_{ISC} .⁷ Since the calculations have been performed in vacuo, without the explicit consideration of solvent effects, the answer provided here can be regarded as a characteristic molecular property of the nucleobases, which might be expected to be somewhat disturbed by the specific environment in solution, in a solid, in vitro, or in vivo. The presence of the same ultrafast decays has been, however, identified in strands of oligonucleotides in solution,⁴⁹ probably related with the channels of the monomers in relatively unstacked nucleobases.⁵⁰

II. Methods and Computational Details

The present calculations include CASSCF geometry optimizations, MEPs, CIs, and STC searches, followed by multiconfigurational perturbation theory, CASPT2, calculations at the optimized geometries. SOC terms and transition dipole moments (TDM) have also been computed. Radiative lifetimes have been estimated by using the Strickler–Berg relationship,⁵¹ as explained elsewhere,⁵² although their applicability is restricted to cases where radiative deactivation predominates. Their magnitude, otherwise, is only indicative of the prospective emissive characteristics of the state related with the TDM values. For the sake of consistency with previous calculations on the singlet states of the systems, the same one-electron basis sets and active spaces were employed. For the pyrimidine T, U, and C and purine A and G nucleobases, basis sets of the ANO-S type contracted to C,N,O[3s2p1d]/H[2s1p] and 6-31G(d,p) were used, respectively. The final results can be described as CASPT2-(14,10) for T, U, and C, involving an active space of 14 electrons distributed in 10 orbitals, with all valence $\pi\pi^*$ and lone-pair orbitals, and CASPT2(14,12) for A and G, which include all $\pi\pi^*$ orbitals except those related to the deepest canonical orbital plus two lone-pair orbitals. Other active spaces were employed in the optimization procedures,

Table 1. Computed Properties for the Low-Lying Singlet and Triplet Excited States of Adenine

State	vertical transition (eV)		band origin (T _e , eV)		τ_{rad}^b
	CASSCF	CASPT2 ^a	CASSCF	CASPT2	
$^1(n\pi^*)$	5.95	4.96 (0.004)	4.88	4.52	334 ns
$^1(L_b \pi\pi^*)$	5.56	5.16 (0.004)	4.92	4.83	251 ns
$^1(L_a \pi\pi^*)^c$	7.03	5.35 (0.175)			
$^3(L_a \pi\pi^*)$	3.77	4.00	3.52	3.36 ^d	359 ms
$^3(n\pi^*)$	5.38	4.91	4.84	4.41	
$^3(\pi\pi^*)$	5.07	4.95			

^a Oscillator strengths within parentheses. ^b Computed using the Strickler–Berg approximation. See SI. ^c Geometry optimization leads directly to a conical intersection with the ground state, (gs/ $\pi\pi^*$)CI, at 4.0 eV. See refs 29 and 38. ^d Phosphorescence band origin and maximum in solution/glasses: 3.43 and 3.05 eV, respectively. See refs 58 and 59.

following a strategy which was proved successful previously. More detailed technical aspects of the calculations can be found in our previous papers^{23–25,37–39} and in the Supporting Information (SI). All the reported calculations used the quantum-chemical methods implemented in the MOLCAS 7 package.^{53,54}

III. Results and Discussion

The research effort in our group has been focused in recent years on the main singlet decay channels involving DNA/RNA nucleobases as well as several derivatives.^{18,20,37–39} In addition, studies were reported on the lowest triplet population mechanisms of the pyrimidine nucleobases thymine,^{24,26} uracil,^{25,26} and cytosine.²³ Other recent theoretical studies on the vertical and adiabatic energies of the nucleobases' triplet states have also been reported.^{55,56} In the present paper, we outline a unified scheme describing prospective population paths of the triplet manifold in all five natural DNA/RNA nucleobases T, U, C, A, and G, in order to obtain an overall model able to explain the common and the distinct behavior of the systems. Fully new results on the triplet states of the purine nucleobases A and G shall be presented, whereas our previous studies on T and U and new complementary calculations on C will be used and commented upon. The following subsections describe the results for each of the nucleobases. The most relevant conclusions are summarized in the last section.

A. Population of the Triplet Manifold in Adenine.

Table 1 compiles vertical transitions, band origins, oscillator strengths, and radiative lifetimes computed for the transitions to the singlet and triplet states of adenine at the CASSCF and CASPT2 levels of theory. Unless indicated, CASPT2 results will be used in the discussion. Both at the FC region and adiabatically, the lowest-lying singlet excited state is of the $n\pi^*$ ($n_N\pi^*$) type, whereas the one carrying the largest intensity for the related transition, and therefore getting initially most of the population at low energies almost up to 6.0 eV, is the $^1(\pi\pi^*)$ HOMO (H) \rightarrow LUMO (L) (hereafter L_a) singlet excited state at 5.35 eV. The ultrafast nonradiative decay undergone by adenine in the femtosecond range^{17,47} can be rationalized by the barrierless character of the path on this state leading from the FC region toward a CI seam with the ground state, (gs/ $^1\pi\pi^*$)CI,^{29,38,41,43,57} and it is shown also here in Figure 2. Unlike simple geometry optimizations,

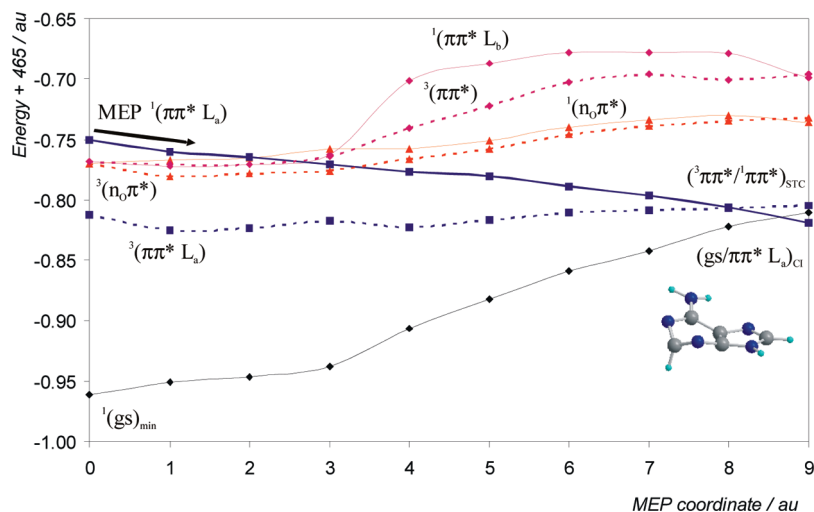


Figure 2. Evolution of the ground and lowest singlet and triplet excited states for adenine from the FC geometry along the $^1(\pi\pi^* L_a)$ MEP.

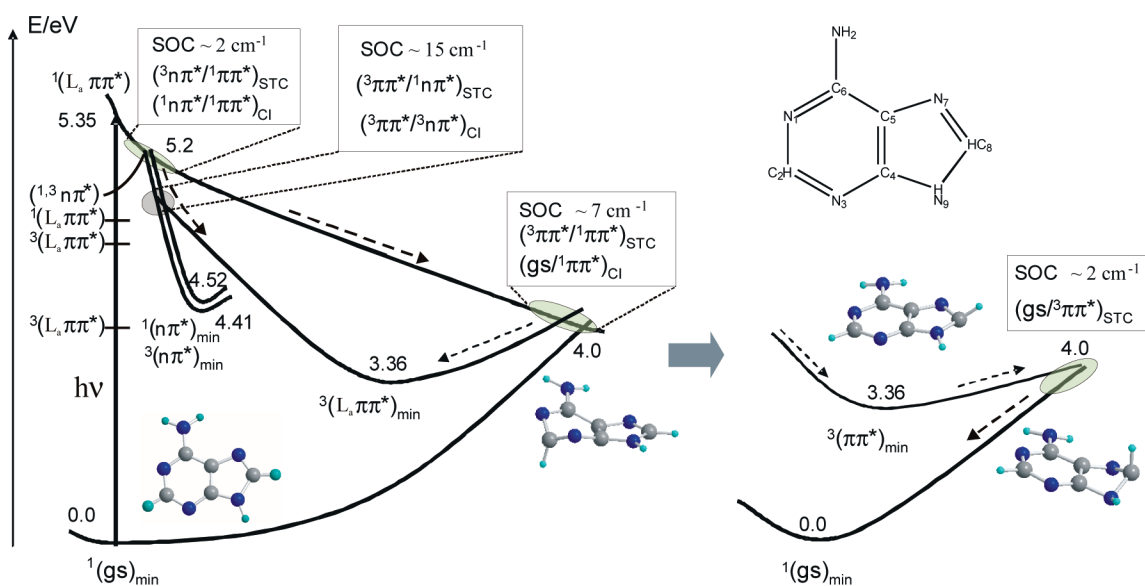


Figure 3. Scheme, based on CASPT2 results, of the photochemistry of adenine focused on the population of the lowest-energy triplet state. Unless otherwise stated, $^1\pi\pi^*$ represents the $^1L_a \pi\pi^*$ state.

the use of the MEP technique guarantees the absence of energy barriers along the lowest-energy path. The structure of the CI at the end of the MEP can be characterized as methanamine-like, involving combined stretching and twisting of the $C_2=N_3$ bond (analogous to an ethene-like CI).^{37–39,60} The presence of an accessible CI explains also the low fluorescence quantum yield ($\sim\phi_F = 10^{-4}$) detected for adenine with a band origin near 4.4 eV in water.¹⁷ This weak emission can be related to the presence of a more polar 7H isomer in solution.^{29,38} A nonfluorescent $^1(n\pi^*)$ minimum is found at 4.52 eV (see Table 1) with a minor contribution to the emissive properties. Similar vertical and adiabatic energy values have been found at other levels of theory.^{21,40,41,43–45,61}

Triplet $\pi\pi^*$ -type states typically lie much lower in energy (here, the lowest one is placed near 1.3 eV) than their singlet counterparts, unlike for $n\pi^*$ -type states, in which a small exchange integral term leads the triplet to be just slightly below the corresponding singlet state. In adenine, for instance, the lowest-energy $^3(n\pi^*)$ state lies, both vertically

and adiabatically, less than 0.1 eV below its singlet analogous state. The consequences for the triplet photophysics of the system are important. Direct singlet $^1(\pi\pi^*)$ –triplet $^3(\pi\pi^*)$ energy transfer seems unlikely in the FC region, where the molecule is almost planar, because of both the large energy gap and low electronic SOC terms ($<0.1 \text{ cm}^{-1}$). The presence of two almost degenerate singlet and triplet $n\pi^*$ -type states at the ground-state geometry can be, however, of high relevance. Along the main decay pathway on S_1 , $^1(\pi\pi^* L_a)$, the state becomes degenerate with different triplet states. As it can be seen in Figures 2 and 3, along the $^1(\pi\pi^* L_a)$ state MEP, two singlet–triplet crossings are described: one at 5.2 eV with the $^3(n\pi^*)$ triplet state, $(^3n\pi^*/^1\pi\pi^*)_{STC}$, and another at 4.0 eV, further along the relaxation path and near the methanamine-like CI with the ground state. The latter crossing involves directly the lowest $^3(\pi\pi^*)$ T₁ triplet state, $(^3\pi\pi^*/^1\pi\pi^*)_{STC}$, and it has a structure displaying the same type of envelope puckered geometry³⁹ with a stretched and twisted double $C_2=N_3$ bond, as at the $(gs/^1\pi\pi^*)$ CI.^{29,38} At

these two STC regions, the computed electronic SOC terms are 2 (${}^3n\pi^*/{}^1\pi\pi^*$) and 7 cm^{-1} (${}^3\pi\pi^*/{}^1\pi\pi^*$). These values can be considered in agreement with the qualitative El-Sayed rules, which pointed to large SOC terms for states of different natures and small otherwise.⁶² El-Sayed rules were developed for molecules near the FC region, where most of the (organic) molecules considered were planar, and their identity, $\pi\pi^*$, $n\pi^*$, etc., could be qualitatively described as such. Far from the FC region, in particular, close to a strongly distorted and puckered geometry like the ${}^2\text{E CI}$, the same rules are not so easy to apply. For instance, the $\pi\pi^*$ state at this region, due to the out-of-plane distortion, has a close diradical character with two electrons in orbitals that are almost perpendicular to each other, the same as the $n\pi^*$ state in the FC region. This effect is particularly true for the low-energy singlet–triplet crossing region, which will be shown to be common in all nucleobases. The presence of the STCs combined with large electronic SOC terms are necessary, but not sufficient, conditions to guarantee efficient ISC processes, but they are good indications of relevant regions in which the population transfer toward the triplet states may take place, provided that the wave packet remains there for a long enough time for the ISC process to take place. The high-energy (~ 5.2 eV) ${}^1\pi\pi^* - {}^3n\pi^*$ STC area, not far from the FC absorption region, fulfills those conditions. On the other hand, recent reaction dynamics calculations suggest⁴⁵ that the region of the (gs/ ${}^1\pi\pi^*$)_{CI} (reached in femtoseconds), where also the STC takes place, represents an area in which the system stays trapped for some time (due to the structure of the CI) until the population switch toward the ground state takes place, which could also explain the slower picosecond channel observed in nucleobases.¹⁷ Figure 3 includes a scheme describing the population of T_1 based on our CASPT2 calculations.

From each one of the STC regions, we have computed corresponding MEPs along the populated triplet states, ${}^3(n\pi^*)$ and ${}^3(\pi\pi^*)$, for the suggested high- and low-energy ISC channels, respectively (they can be found in the SI). Soon, along the MEP on ${}^3(n\pi^*)$, a crossing with the lowest-lying ${}^3(\pi\pi^*)$ state takes place. The corresponding CI, (${}^3n\pi^*/{}^3\pi\pi^*$)_{CI}, represents another funnel for efficient energy transfer within the triplet manifold. Additionally, as the singlet ${}^1(n\pi^*)$ state lies very close to the triplet counterpart and their PEHs run almost parallel, an STC (${}^1n\pi^*/{}^3\pi\pi^*$) also occurs at that region. Considering that the computed SOC term in this case rises to 15 cm^{-1} , the corresponding ISC process toward the ${}^3(\pi\pi^*)$ state should be considered very favorable. A subsequent MEP from the (${}^3n\pi^*/{}^3\pi\pi^*$)_{CI} along the ${}^3(\pi\pi^*)$ PEH leads to the lowest triplet state minimum (see SI). Regarding the STC described at 4.0 eV, the MEP computed from the (${}^3\pi\pi^*/{}^1\pi\pi^*$)_{STC} along the ${}^3(\pi\pi^*)$ state leads directly to the minimum of the triplet state (see SI). The involvement of a dark singlet $n\pi^*$ state on adenine relaxation dynamics was previously suggested by other authors to explain slow decay features.^{17,63,64}

After the lowest triplet state is populated by any of the previous ISC processes, the system is finally expected to evolve toward the triplet state minimum, ${}^3(\pi\pi^*)_{\text{min}}$ (see Figure 3), which is characterized by a structure with almost planar rings but with the terminal hydrogen C_8H lifted near

40° and with an increased bond length C_2N_3 of 1.389 Å (compared to 1.311 Å in the ground state), in agreement with previous estimations.⁵⁵ The reactivity that could be attributed to this triplet state originates from its biradical character on C_2 and N_3 . The minimum is placed at 3.36 eV adiabatically (see Table 1) from the ground state optimized minimum, a value consistent with the measured phosphorescence band origin in solution at 3.43 ,⁵⁸ and other theoretical results.^{43,55} We have also located the singlet–triplet crossing connecting the ${}^3(\pi\pi^*)$ and the ground state and mapped the MEP leading from such an STC toward ${}^3(\pi\pi^*)_{\text{min}}$ (see SI). The crossing is placed near 4.0 eV from the ground state minimum, which means that there is a barrier of near 0.6 eV (14.0 kcal/mol) to reach (gs/ ${}^3\pi\pi^*$)_{STC} from ${}^3(\pi\pi^*)_{\text{min}}$. The distortion of the five-membered ring is larger at the STC point, and the computed electronic SOC is somewhat low, ~ 2 cm^{-1} , suggesting for the triplet state a long lifetime and a slow relaxation, becoming therefore prone to reacting or transferring its energy by photosensitization mechanisms.^{8–10}

In summary, we have identified in adenine (see Figure 3) three possible intrinsic ISC channels toward the lowest triplet state which can be easily accessed from the main barrierless MEP for singlet decay dynamics, two of them mediated by $n\pi^*$ states. In all three cases, the magnitude of the computed SOC terms between the relevant states is high enough to suggest an efficient population of the triplet manifold in adenine upon UV irradiation. This type of ${}^1\pi\pi^*/{}^3\pi\pi^*$ ISC mechanism via intermediate $n\pi^*$ states can be suggested here as favorable, even far from the FC region, as it has been recently reported also for other biological chromophores such as isoalloxazine⁶⁵ and psoralen.⁶⁶ Both mechanisms described here can in any case contribute to the overall population of the lowest triplet state. In principle, in different environments, such as in polar solvents, it is expected that the $n\pi^*$ -type excited state will become destabilized with respect to $\pi\pi^*$ -type excited states.⁶⁷ Despite those effects, both singlet and triplet $n\pi^*$ -type states are estimated to lie in the solvent below the ${}^1(\pi\pi^* L_a)$ state at the FC geometry,⁶⁸ guaranteeing the existence of the STC crossing upon decay along the ${}^1(\pi\pi^* L_a)$ state. Intersystem crossing quantum yields have been measured by means of nanosecond laser photolysis in adenine to be 0.23×10^{-2} higher than in guanine.⁷ Likewise, phosphorescence quantum yields of 4.5×10^{-2} for adenine in frozen solutions at 77 K have been reported, slightly higher than for guanine, 3.6×10^{-2} and 2×10^{-2} ,^{69,70} and lower than thymine.⁷ For the purine nucleobases, the ISC yield has been measured to be lower in the nucleotide.⁷ Also in adenine,⁷¹ although less clearly documented as in pyrimidine nucleobases, a wavelength dependence of the intersystem crossing quantum yield in nucleobases has been reported, as it can be expected by the contribution of the three (at excitation energies higher than 5.0 eV) or just the lowest-energy (at energies close to 4.0 eV) ISC mechanisms. This point requires further experimental confirmation.

B. Population of the Triplet Manifold in Guanine. The same strategy as for adenine has been followed in the calculations of guanine. Table 2 lists the main spectroscopic properties of the lowest-lying singlet and triplet states of the

Table 2. Computed Properties for the Low-Lying Singlet and Triplet Excited States of Guanine

state	vertical transition (eV)		band origin (T _e , eV)		τ _{rad} ^b
	CASSCF	CASPT2 ^a	CASSCF	CASPT2	
¹ (L _a ππ*) ^c	6.36	4.93 (0.158)			
¹ (n _O π*)	5.70	5.54 (0.002)	4.04	4.56	6800 ns
¹ (L _b ππ*)	7.04	5.72 (0.145)	6.07	5.69	5 ns
³ (L _a ππ*)	3.97	4.11	3.13	3.15	3562 ms
³ (ππ*)	5.08	4.76			
³ (ππ*)	5.41	5.14			
³ (n _O π*)	5.82	5.30	4.64	4.17	

^a Oscillator strengths within parentheses. ^b Computed using the Strickler–Berg approximation. See SI. ^c Geometry optimization leads directly to a conical intersection with the ground state, (gs/ππ*)CI, at 4.3 eV. See ref.³⁹

molecule. As compared to adenine, a couple of important aspects of the electronic structure of guanine have to be highlighted. First of all is the low energy displayed by the ¹(ππ* L_a) HOMO → LUMO state, placed at 4.93 eV at the FC region as the lowest-energy feature. The value of the related oscillator strength, 0.158, indicates that this is the bright singlet state basically populated in the low-energy absorption spectrum, and that the relevant photophysics of the system will take place along the MEP on such a state. The second aspect is related to the high-energy of the low-lying nπ* states, which are placed near 0.6 (singlet) and 0.4 (triplet) eV above the ¹(ππ* L_a) state (even higher in solution). As is clear from Table 2, and also from Figure 4, the gap between the initially populated singlet state and the nπ* states is much larger than in adenine. At the FC region, it is therefore expected that an ISC process relating the ¹(ππ* L_a) and ³nπ* states is less favorable than for adenine.

Figure 4 displays the MEP from the FC structure and along the ¹(ππ* L_a) state. At the beginning of the MEP, the singlet state only crosses with the second triplet ³(ππ*) state. The computed electronic SOC terms are small (<0.1 cm⁻¹), and only strongly coupled vibronic terms would enhance in this region the ISC rate. Near point 9 of the MEP, the singlet state crosses with the lowest triplet state, as it occurred in adenine. The STC region, placed adiabatically at 4.3 eV, is not far from the CI between the singlet and the ground state. The corresponding SOC terms are much larger here, 8 cm⁻¹, and therefore a more efficient ISC process leading directly to the population of the lowest ³(ππ*) state can be therefore expected, or at least proposed. As compared with adenine, however, the overall population of the triplet manifold cannot be expected to be favorable. Even when the ³(n_Oπ*) excited state minimum lies lower in energy than the (gs/ππ*)_{CI}, and therefore a crossing with the ¹(ππ*) state takes place at some other region, the key point is that such a crossing cannot be easily accessed from the photochemically relevant MEP, that is, the main decay path for singlet deactivation. As a matter of fact, we have computed the STC crossing structure (³n_Oπ*/¹ππ*)_{STC}, which lies almost degenerate with the computed (¹n_Oπ*/¹ππ*)_{CI} (see ref 39), at 4.6 eV, but far from the main MEP region, because it represents the stretching and twisting of the C₆N₁ bond. Even when such a structure, in which the SOC is large enough, 8 cm⁻¹, can be accessed with excess energy, it cannot be considered as favorable as those reached via the MEP-related channels.

For the sake of completeness, we have connected the mentioned STC points with the minimum of the lowest ³(ππ*) state by computing the corresponding MEPs: (i) from the computed (³ππ*/¹ππ*)_{STC} and (³n_Oπ*/¹ππ*)_{STC} structures along the ³(ππ*) and ³(n_Oπ*) states, leading to their respective minima, (ii) from the computed (³n_Oπ*/³ππ*)_{CI} to the ³(ππ*) minimum, and (iii) from the singlet–triplet (¹gs/³ππ*)_{STC} toward the final ³(ππ*) minimum. All them are possible paths leading to the population of the lowest triplet state, although we emphasize that, unlike adenine, only the lowest-lying 4.3 eV ISC mechanism related to the (³ππ*/¹ππ*)_{STC} should be initially considered efficient, because it is the only one taking place in the proximity of the main ¹(ππ* L_a) MEP (see Figure 5 for a scheme of the triplet photophysics in guanine). Finally, the ³(ππ*) minimum has been connected through a corresponding MEP with the STC with the ground state, (gs/³ππ*)_{STC}. Although the SOC terms at this point are higher than in adenine, the barrier from the minimum, placed at 3.15 eV, is too large (0.85 eV) to expect an efficient decay to the ground state. All computed MEPs can be found in the SI. At the ³(ππ*) minimum, the molecule displays a slightly puckered envelope structure on the six-membered ring,³⁹ with the C₂N₃ bond having a biradical character and enlarged up to 1.438 Å, as compared to 1.286 Å at the FC ground-state geometry.

It has to be finally mentioned that guanine is the only natural nucleobase in which no phosphorescence data or triplet state formation has been reported for the parent compound, although intersystem crossing⁷ and phosphorescence quantum yields of 0.042 and 0.095 have been reported for the nucleoside and nucleotide in ethanol,⁶ 5 to 7 times larger than the fluorescence quantum yields. It has to be remembered also that the natural keto form of 9H-guanine is not the most stable in the gas phase and that other close tautomers can contribute to the measurements for the isolated system,^{39,72} not in an oligomer sample.

C. Population of the Triplet Manifold in Pyrimidine Nucleobases: Thymine, Uracil, and Cytosine. For the sake of brevity, we will discuss the triplet manifold population of the pyrimidine nucleobases together within the same framework. The computational strategies followed have been those described above for adenine and guanine. As uracil has a state structure and triplet photophysics very similar to that of thymine, we will refer to our previous results²⁵ and concentrate on the latter. Thymine has, at the FC region, a low-lying ¹(n_Oπ*) state (basically related to the O₄ atom), placed 0.2 eV below the spectroscopic ¹(ππ* L_a) HOMO → LUMO state, this one lying at 4.89 eV with a related oscillator strength of 0.167 (see Table 3). The photophysical mechanisms proposed for the population of the lowest triplet state will be very similar to those already explained for adenine, as Table 3 and Figures 6 and 7 can confirm. Once more, the key point is that three different STC regions can be easily accessed through the main decay path of the energy, as it is the FC ¹(ππ* L_a) MEP, being prospective channels for ISC toward the lowest-lying triplet state.

Soon after the beginning of the MEP (see Figure 6), the ¹(ππ* L_a) state crosses with both singlet and triplet nπ* states. Apart from a possible IC toward the singlet ¹(n_Oπ*)

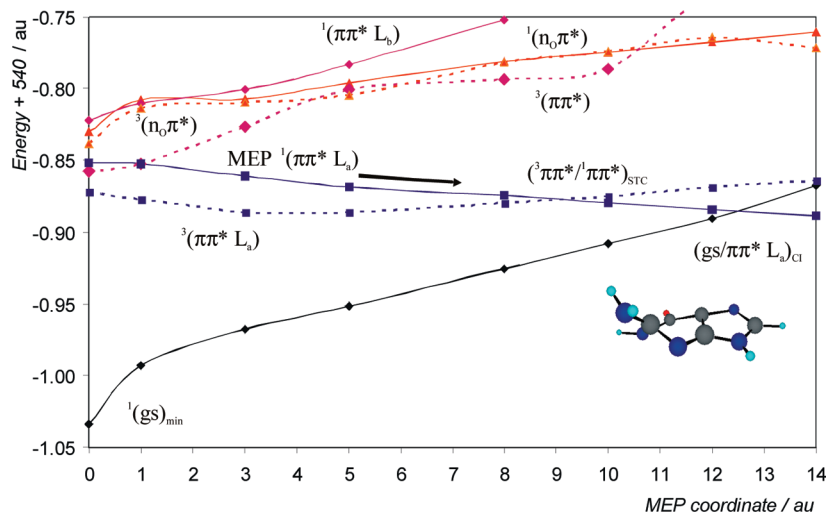


Figure 4. Evolution of the ground and lowest singlet and triplet excited states for guanine from the FC geometry along the $^1(\pi\pi^* L_a)$ MEP.

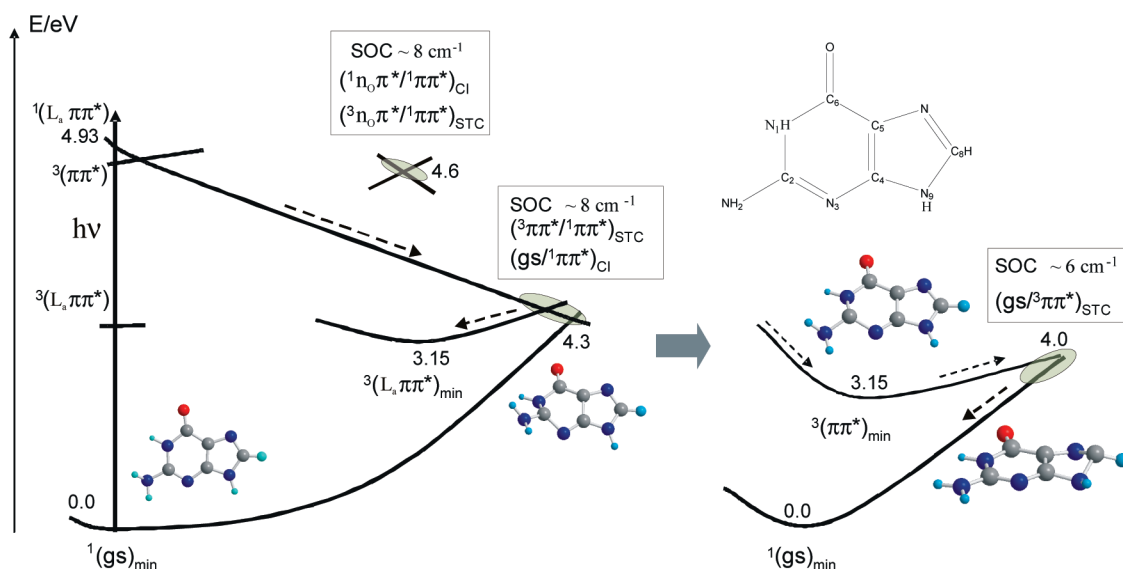


Figure 5. Scheme, based on CASPT2 results, of the photochemistry of guanine focused on the population of the lowest-energy triplet state. Unless otherwise stated, $^1\pi\pi^*$ represents the $^1L_a \pi\pi^*$ state.

Table 3. Computed Properties for the Low-Lying Singlet and Triplet Excited States of Thymine^a

state	vertical transition (eV)		band origin (T_e , eV)		τ_{rad}^c
	CASSCF	CASPT2 ^b	CASSCF	CASPT2	
$^1(nO\pi^*)^d$	5.41	4.77 (0.004)	4.23	4.05	2501 ns
$^1(\pi\pi^* L_a)$	6.52	4.89 (0.167)	6.07	4.49	9 ns
$^1(\pi\pi^*)$	7.36	5.94 (0.114)			
$^3(\pi\pi^* L_a)$	3.95	3.59	2.99	2.87	17 ms
$^3(nO\pi^*)^d$	5.21	4.75	3.84	3.93	
$^3(\pi\pi^*)$	5.86	5.14			

^a See also ref 24. ^b Oscillator strengths within parentheses.

^c Computed using the Strickler–Berg approximation. See SI.

^d Involving basically O₄ (in ortho position with methyl group).

state through a corresponding CI, this region may be responsible for the first ISC process taking place in thymine at high energies (4.8 eV), in which the $^3(nO\pi^*)$ state could be populated from the initially activated singlet $\pi\pi^*$ state. The SOC terms, computed as 8 cm^{-1} , point to the efficiency of the process. Another MEP computed from this crossing and along the $^3(nO\pi^*)$ PEH leads the system toward the

minimum of this state, in whose neighborhood we have found the conical intersection with the lowest triplet state, ($^3nO\pi^*/^3\pi\pi^*$)_{CI}, near 3.9 eV. As the singlet and triplet $n\pi^*$ PEHs are always very close along the MEP, near the CI we have also found the ($^1nO\pi^*/^3\pi\pi^*$)_{STC}. In case some population reaches the $^1(nO\pi^*)$ state via the higher-energy crossing with $^1(\pi\pi^* L_a)$ —and a decay path through this dark intermediate has been recently reported⁶⁸—the energy switch toward the lowest triplet state would be extremely favorable, because the computed SOC term increases in the ($^1nO\pi^*/^3\pi\pi^*$)_{STC} region to 61 cm^{-1} . It is possible to confirm our suggestions about the effectiveness of this type of mechanism thanks to the recent study by Etinski et al.,²⁶ which has established the efficiency of the ($^1nO\pi^*/^3\pi\pi^*$)_{STC} ISC channel by computing vibrational FC factors and ISC rates. Either by triplet–triplet IC or by singlet–triplet ISC, the final population process of the lowest $^3(\pi\pi^*)$ state should be considered to be extremely favorable. As in the other nucleobases, a low-energy STC region lies close to the end of the FC $^1(\pi\pi^*$

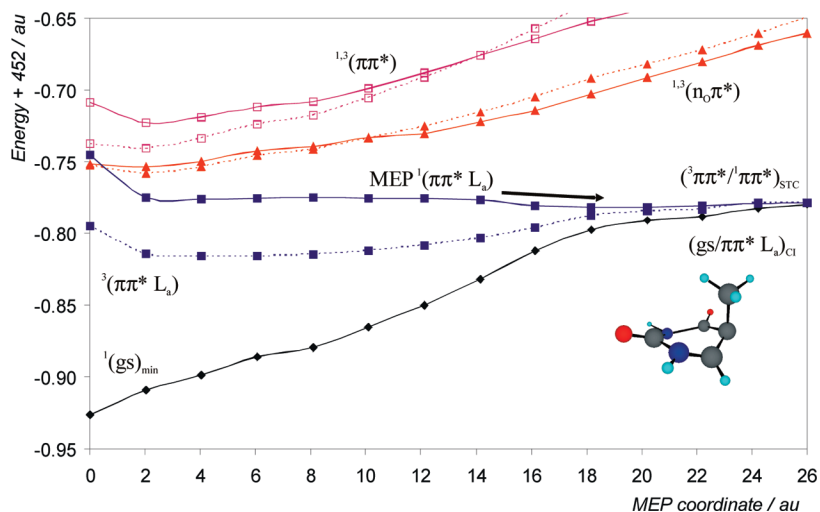


Figure 6. Evolution of the ground and lowest singlet excited states for thymine from the FC geometry along the $^1(\pi\pi^* L_a)$ MEP.

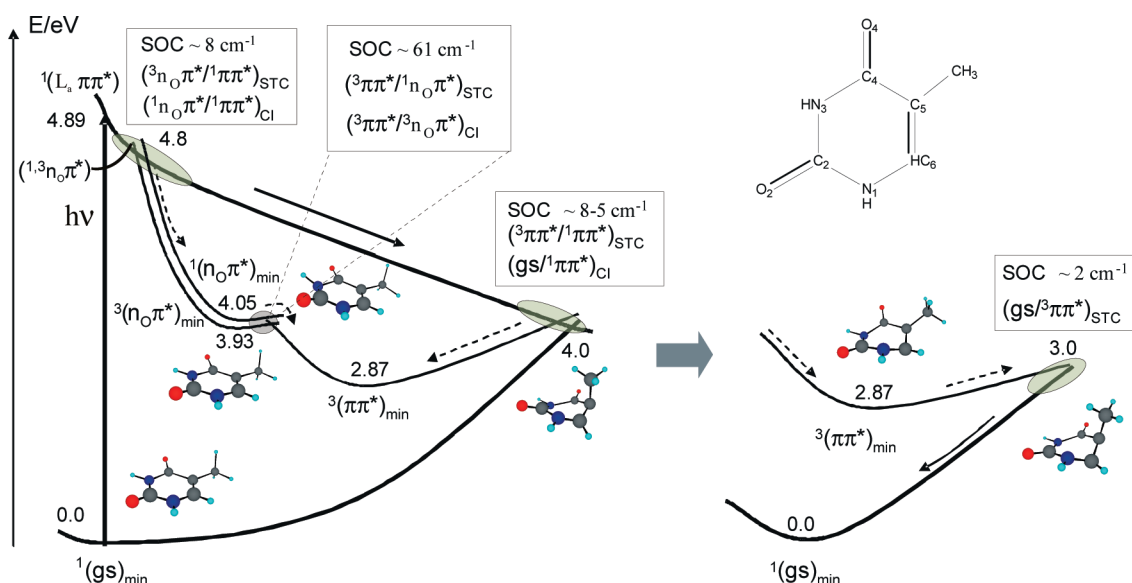


Figure 7. Scheme, based on CASPT2 results, of the photochemistry of thymine focused on the population of the lowest-energy triplet state. Unless otherwise stated, $^1\pi\pi^*$ represents the $^1L_a \pi\pi^*$ state.

L_a MEP, at 4.0 eV. As observed in Figure 6, this area of the PEH is flat and extended close to the end of the MEP. The SOC values computed at different points along the path range from 5 to 8 cm^{-1} . The efficiency of the process would be also high if, as in adenine, the wave packet decaying through the singlet manifold is delayed in the region of the singlet–singlet CI. The present model allows for an understanding of the reported wavelength dependence on the ISC quantum yield in nucleobases, surely caused by the location of the two STC interacting regions and their accessibility upon the initial excitation conditions. In the case of thymine, the value increases from 3.9×10^{-3} at 280 nm (4.43 eV), where only the lowest-energy channel can be reached, to 5.2×10^{-2} at 240 nm (5.17 eV),^{7,73} where both described channels are accessible.

As for the purine nucleobases, MEPs connecting the different critical points have been computed (see SI). The lowest triplet state may be populated by any of the previous ISC processes. At the state minimum, the molecule displays

a distorted structure with a ring deformation including the dihedral angle $C_2N_1C_6C_5$ as 44° and an increased bond length C_5C_6 of 1.494 \AA with certain biradical character. The minimum is placed at 2.87 eV adiabatically from the ground state optimized minimum, a value somewhat lower than the 3.2 eV estimated for the location of the triplet state for the thymine mononucleotide in aqueous solution at room temperature⁹ and consistent with previous theoretical determinations at around 2.8–3.0 eV.⁷⁴ As a final aspect of the evolution along the triplet manifold in thymine, we have located the singlet–triplet crossing connecting the $^3(\pi\pi^*)$ and the ground state and mapped the MEP leading from such an STC toward $^3(\pi\pi^*)_{\text{min}}$ (see SI). The crossing is placed near 3.0 eV from the ground state minimum, which means that there is a barrier of 0.13 eV (3.0 kcal/mol) to reach $(\text{gs}/^3\pi\pi^*)_{\text{STC}}$ from $^3(\pi\pi^*)_{\text{min}}$, and the molecule recovers there the planarity. Although the computed electronic SOC is somewhat low, $\sim 2 \text{ cm}^{-1}$, a barrier which is smaller than that for purines may explain the shorter triplet lifetimes

Table 4. Computed Properties for the Low-Lying Singlet and Triplet Excited States of Cytosine

state	vertical transition (eV)		band origin (T_e , eV)		τ_{rad}^b
	CASSCF	CASPT2 ^a	CASSCF	CASPT2	
¹ (L _a $\pi\pi^*$) ^c	5.22	4.41 (0.069)	4.14	3.62	30 ns
¹ (n _O π^*)	5.23	4.95 (0.001)	3.68	3.72	1200 ns
¹ (n _N π^*) ^d	5.59	5.06 (0.003)			
¹ (L _b $\pi\pi^*$)	6.17	5.89 (0.106)			
³ (L _a $\pi\pi^*$)	3.64	3.53	2.85	2.98	437 ms
³ ($\pi\pi^*$)	4.87	4.45			
³ (n _O π^*)	5.13	4.63	3.49	3.66	
³ (n _N π^*)	5.31	4.94			

^a Oscillator strengths within parentheses. ^b Computed using the Strickler–Berg approximation. See SI. ^c The MEP to the minimum and the CI, (gs/ $\pi\pi^*$)_{CI}, at 3.6 eV, are competitive. See ref 37. ^d Geometry optimization leads directly to a CI with the ground state, (gs/n_N π^*)_{CI}. See ref 76.

measured for pyrimidine (~ 0.6 s) than for purine (~ 2.0 s) nucleobases in ethanol glasses.⁶ Similar conclusions can be derived for uracil, which has a state structure and properties very similar to those of thymine.^{25,26}

Regarding cytosine, the values in Table 4 help to understand (and predict to some extent) the behavior of its triplet photophysics. As in guanine, cytosine has a lowest-lying singlet ¹($\pi\pi^*$ L_a) state, whose initial interaction with the $n\pi^*$ states, placed higher in energy, will not be strong either at the FC region or along the ¹($\pi\pi^*$ L_a) decay pathway (see Figure 8). The singlet relaxation in C is somewhat more complex than in the other nucleobases. The presence of a low-lying planar minimum for the ¹($\pi\pi^*$ L_a) state at 3.62 eV, nearly isoenergetic with the ethene-like (gs/ $\pi\pi^*$)_{CI}, generates several competitive decay paths, as has been analyzed before.^{23,37,75} The possibilities for displaying different ISC processes are therefore larger, but always at low, not at high, energies like, for instance, in thymine, uracil, or adenine. In particular, we show in Figure 8 a linear interpolation in internal coordinates (LIIC) path from the FC region toward the ethene-like CI with the ground state. The barrier along the ¹($\pi\pi^*$ L_a) state, computed 2.5 kcal mol⁻¹ as a higher bound, is very small, and in practice the path can be considered barrierless. As in the other nucleobases, an STC between the lowest $\pi\pi^*$ states takes place close to the CI, at 3.6 eV, yielding a SOC term value of 6 cm⁻¹. In a previous study,²³ we analyzed ISC processes taking place at other low-energy regions, obtaining also large SOC values and expectedly favorable situations for the lowest triplet population.

As a result of the excited state structure in C, obtained at the CASPT2 level, the photophysical scheme for the population of the lowest triplet state of the molecule can be summarized in Figure 9. Unlike in the other two pyrimidine nucleobases, where three basic channels for the possible triplet manifold population were found, one at high energies (close to FC and $n\pi^*$ mediated) and another at low energies (caused by the common ethene-like CI type of decay present in all nucleobases), in C, only low-energy channels seem to be accessible. This feature could probably help to explain the absence of cytosine (guanine too) components in DNA phosphorescence at low temperatures,^{1–3} and also the generally lower phosphorescence quantum yields obtained for cytosine and its derivatives as compared to other

nucleobases.⁶ The same trends are obtained for ISC yields from flash photolysis experiments in nucleotides, although not in nucleobases.⁷ Higher yields of $n\pi^*$ formation have been suggested for cytosine than thymine,⁶⁸ but theoretical evidence indicates that the higher-lying $n\pi^*$ states of cytosine will be less accessible from the main relaxation pathways than in thymine due to the large potential energy barriers found in the former.⁷⁶

IV. Summary and Conclusions

Calculation of PEHs for the low-lying singlet and triplet states of natural DNA/RNA nucleobases adenine, guanine, thymine, uracil, and cytosine at the *ab initio* multiconfigurational CASPT2//CASSCF quantum-chemical level have been carried out in order to help to establish general mechanisms for the population of the triplet manifold of the molecules. The proposed framework is an attempt to rationalize the reported triplet states properties of DNA components, in particular the measurement of larger quantum yields of phosphorescence than of fluorescence in the individual systems,^{4,7} the observed wavelength dependence of the triplet state formation,^{7,73} or the prevalence of adenine and thymine components in the phosphorescence signals of DNA at low temperatures.^{1–3} It can be considered that an efficient ISC channel is easily accessible from the regions close to the main decay pathway of the initially populated singlet state. We have analyzed the accessibility of the ISC channels for the population of the lowest triplet state along such a pathway, a strategy that requires computation of minimum energy paths on the different states and determination of singlet–triplet crossings and conical intersections. This is, however, only a necessary but not sufficient condition to establish the efficiency of an ISC process. Computation of vibronic contributions to the ISC rates and reaction dynamic calculations establishing the temporal evolution of the system are encouraged in a close future in order to unambiguously determine if the proposed accessible singlet–triplet crossing regions fulfill all the requirements: close singlet–triplet energies, a high density of vibronic states, large vibronic contributions to the spin–orbit coupling terms, and regions where the population gets trapped for long enough of a time to allow the ISC process to take place in competition with the internal conversion decay, for instance, close to the FC region, to a singlet state minimum, or near a sloped conical intersection. Recent ISC rate calculations on thymine and uracil²⁶ confirm the main role of some of our proposed ISC mechanisms in these systems.

Our results indicate that three STC regions can be easily accessed from the singlet main decay pathway in adenine, thymine, and uracil, two of them located at high energies and mediated by the presence of lowest-lying singlet and triplet $n\pi^*$ states, and a third one at low energies close to the end of the main MEP on the ¹($\pi\pi^*$) singlet excited state and the ethene-like (pyrimidines) or methanamine-like (purines) conical intersection of this state with the ground state. These three regions are proposed as prospective ISC channels. At least those related to the ¹ $n\pi^*$ –³ $\pi\pi^*$ STC seem to be confirmed as such by recent calculations on ISC rates on pyrimidine nucleobases.²⁶ Additionally, the wavelength

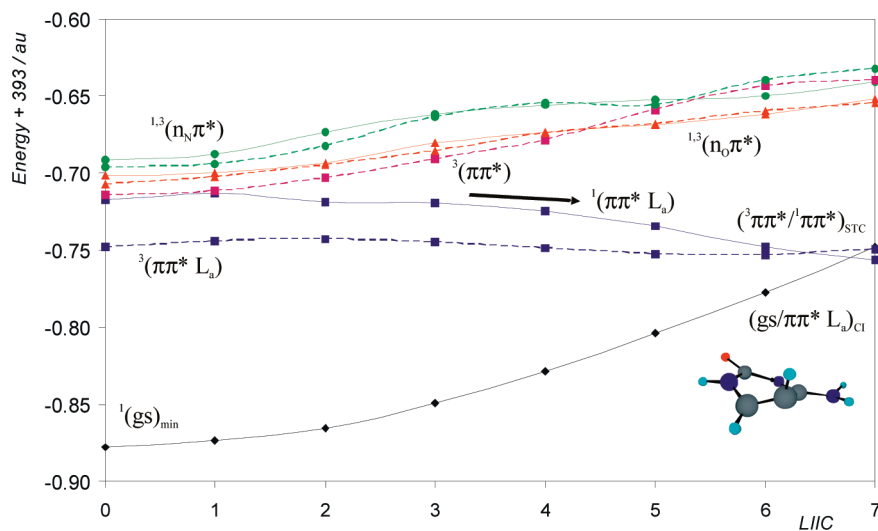


Figure 8. Evolution of the ground and lowest singlet and triplet excited states of cytosine from the FC geometry to the $(gs/\pi\pi^* L_a)_{Cl}$ along a LIIC path competitive with the ${}^1(\pi\pi^* L_a)$ MEP.

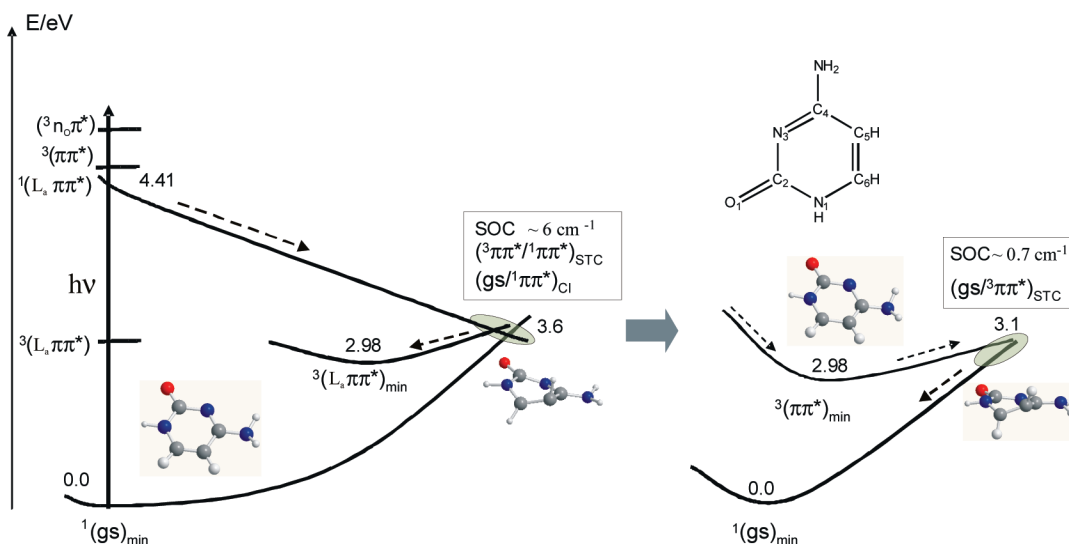


Figure 9. Scheme, based on CASPT2 results, of the photochemistry of cytosine focused on the population of the lowest-energy triplet state.

dependence of the triplet formation quantum yield reported in these three molecules is suggested to be related to the activation of the three (both at high and low excitation energies) or only one (at low energies) ISC channels. On the other hand, guanine and cytosine, having a much lower spectroscopic ${}^1(\pi\pi^*)$ singlet excited state below the $n\pi^*$ -type states, are not expected to display the $n\pi^*$ -mediated ISC mechanisms in regions close to the main MEP and may have only efficient ISC funnels at low energies, close to the singlet CI, a feature common to all nucleobases. The present results explain the fact that guanine and cytosine contribute much less to the phosphorescence of DNA, as it has been established.^{1–3} It is noteworthy to indicate that the phosphorescence spectrum of RNA was also reported,⁷⁷ and it was shown, first, to be determined mainly by the individual properties of the ribonucleotides' π -electron systems, and second, to be composed by triplet signals of adenosine groups and centers of an unknown nature with structureless long-wavelength phosphorescence different from that in DNA. The present results would indicate that adenine and, in this

case, uracil nucleobases should be preferably considered as sources of phosphorescence in RNA, as adenine and thymine are in DNA. It is clear that the present results for the isolated systems cannot be directly extrapolated to polymeric DNA/RNA. As already explained before, the described properties should be, however, considered intrinsic features of the nucleobases that, even if they may change in condensed phases or, in general, in the biological environment for the single monomers, are expected to maintain their basic characteristics, as occurs for the singlet states properties and it seems also for triplet states.⁷⁷

Acknowledgment. This research was supported by projects CTQ2007-61260, CTQ2010-14892, and CSD2007-0010 Consolider-Ingenio in Molecular Nanoscience of the Spanish MEC/FEDER and UV-EQUIP09-5764 of the Universitat de València.

Supporting Information Available: Additional computational details, reaction paths, and Cartesian coordinates

of the singular points. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Imakubo, K. *J. Phys. Soc. Jpn.* **1968**, *24*, 1124.
- (2) Szerenyi, P.; Dearman, H. H. *Chem. Phys. Lett.* **1972**, *15*, 81.
- (3) Arce, R.; Rodríguez, G. *J. Photochem.* **1986**, *33*, 89.
- (4) Gueron, M.; Eisinger, J.; Lamola, A. A. Excited States of Nucleic Acids. In *Basic Principles in Nucleic Acid Chemistry*; Tso, P. O. P., Ed.; Academic Press: New York, 1974; Vol. 1, pp 311–398.
- (5) Daniels, M. In *Photochemistry and Photobiology of Nucleic Acids*; Wang, S. Y., Ed.; Academic Press: New York, 1976; Vol. 1, pp 23–108.
- (6) Görner, H. *J. Photochem. Photobiol. B: Biol.* **1990**, *5*, 359.
- (7) Cadet, J.; Vigny, P. In *Bioorganic Photochemistry*; Morrison, H., Ed.; John Wiley & Sons: New York, 1990; Vol. 1, pp 1–272.
- (8) Gut, I. G.; Wood, P. D.; Redmond, R. W. *J. Am. Chem. Soc.* **1996**, *118*, 2366.
- (9) Wood, P. D.; Redmond, R. W. *J. Am. Chem. Soc.* **1996**, *118*, 4256.
- (10) Bosca, F.; Lhiaubet-Vallet, V.; Cuquerella, M. C.; Castell, J. V.; Miranda, M. A. *J. Am. Chem. Soc.* **2006**, *128*, 6318.
- (11) Kang, H.; Lee, K. T.; Jung, B.; Ko, Y. J.; Kim, S. K. *J. Am. Chem. Soc.* **2002**, *124*, 12958–12959.
- (12) Klessinger, M. In *Theoretical Organic Chemistry - Theoretical and Computational Chemistry*; Párkányi, C., Ed.; Elsevier: Amsterdam, 1998; p 581.
- (13) Serrano-Pérez, J. J.; Merchán, M.; Serrano-Andrés, L. *J. Phys. Chem. B.* **2008**, *112*, 14002.
- (14) Brown, I. H.; Johns, H. E. *Photochem. Photobiol.* **1968**, *8*, 273.
- (15) Schreier, W. J.; Schrader, T. E.; Koller, F. O.; Gilch, P.; Crespo-Hernández, C. E.; Swaminathan, V. N.; Carell, T.; Zinth, W.; Kohler, B. *Science* **2007**, *315*, 625.
- (16) Roca-Sanjuán, D.; Olaso-González, G.; González-Ramírez, I.; Serrano-Andrés, L.; Merchán, M. *J. Am. Chem. Soc.* **2008**, *130*, 10768.
- (17) Crespo-Hernández, C. E.; Cohen, B.; Hare, P. M.; Kohler, B. *Chem. Rev.* **2004**, *104*, 1977–2019.
- (18) Serrano-Andrés, L.; Merchán, M. In *Radiation Induced Molecular Phenomena in Nucleic Acid: A Comprehensive Theoretical and Experimental Analysis*; Shukla, M. K., Leszczynski, J., Eds.; Springer: The Netherlands, 2008; pp 435–472.
- (19) Middleton, C. T.; De La Harpe, K.; Su, C.; Law, Y. K.; Crespo-Hernández, C. E.; Kohler, B. *Annu. Rev. Phys. Chem.* **2009**, *60*, 217.
- (20) Serrano-Andrés, L.; Merchán, M. *J. Photochem. Photobiol. C: Photochem. Rev.* **2009**, *10*, 21.
- (21) Conti, I.; Garavelli, M.; Orlandi, G. *J. Am. Chem. Soc.* **2009**, *131*, 16108.
- (22) Conti, I.; Altoè, P.; Stenta, M.; Garavelli, M.; Orlandi, G. *Phys. Chem. Chem. Phys.* **2010**, DOI: 10.1039/b926608a (accessed May 4, 2010).
- (23) Merchán, M.; Serrano-Andrés, L.; Robb, M. A.; Blancafort, L. *J. Am. Chem. Soc.* **2005**, *127*, 1820.
- (24) Serrano-Pérez, J. J.; González-Luque, R.; Merchán, M.; Serrano-Andrés, L. *J. Phys. Chem. B.* **2007**, *111*, 11880.
- (25) Climent, T.; González-Luque, R.; Merchán, M.; Serrano-Andrés, L. *Chem. Phys. Lett.* **2007**, *441*, 327.
- (26) Etinski, M.; Fleig, T.; Marian, C. M. *J. Phys. Chem. A* **2009**, *113*, 11809.
- (27) Klessinger, M.; Michl, J. *Excited States and Photochemistry of Organic Molecules*; VCH Publishers, Inc.: New York, 1995.
- (28) *Computational Photochemistry*; Olivucci, Ed.; Elsevier: Amsterdam, 2005.
- (29) Serrano-Andrés, L.; Merchán, M.; Borin, A. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8691.
- (30) Tatchen, J.; Gilka, N.; Marian, C. M. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5209.
- (31) Salzmann, S.; Tatchen, J.; Marian, C. M. *J. Photochem. Photobiol. A* **2008**, *198*, 221.
- (32) Carpenter, B. K. *Chem. Soc. Rev.* **2006**, *35*, 736.
- (33) Harvey, J. N.; Poli, R.; Smith, K. M. *Coord. Chem. Rev.* **2003**, *238–239*, 347.
- (34) González-Navarrete, P.; Coto, P. B.; Polo, V.; Andrés, J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 7189.
- (35) Atchity, G. J.; Xantheas, S. S.; Ruedenberg, K. *J. Chem. Phys.* **1991**, *95*, 1862.
- (36) Ben-Nun, M.; Molnar, F.; Schulten, K.; Martinez, T. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1769.
- (37) Merchán, M.; González-Luque, R.; Climent, T.; Serrano-Andrés, L.; Rodríguez, E.; Reguero, M.; Peláez, D. *J. Phys. Chem. B* **2006**, *110*, 26471.
- (38) Serrano-Andrés, L.; Merchán, M.; Borin, A. C. *Chem.—Eur. J.* **2006**, *12*, 6559.
- (39) Serrano-Andrés, L.; Merchán, M.; Borin, A. C. *J. Am. Chem. Soc.* **2008**, *130*, 2473.
- (40) Blancafort, L. *J. Am. Chem. Soc.* **2006**, *128*, 210.
- (41) Perun, S.; Sobolewski, A. L.; Domcke, W. *J. Am. Chem. Soc.* **2005**, *127*, 6257.
- (42) Perun, S.; Sobolewski, A. L.; Domcke, W. *J. Phys. Chem. A* **2006**, *110*, 13238.
- (43) Marian, C. M. *J. Chem. Phys.* **2005**, *122*, 104314.
- (44) Chen, H.; Li, S. H. *J. Phys. Chem. A.* **2005**, *109*, 8443.
- (45) Barbatti, M.; Lischka, H. *J. Am. Chem. Soc.* **2008**, *130*, 6831.
- (46) Hudock, H. R.; Martinez, T. J. *ChemPhysChem* **2008**, *9*, 2486.
- (47) Canuel, C.; Mons, M.; Pluzzi, F.; Tardivel, B.; Dimicoli, I.; Elhanine, M. *J. Chem. Phys.* **2005**, *122*, 074316.
- (48) Bernardi, F.; Olivucci, M.; Robb, M. A. *Pure Appl. Chem.* **1995**, *67*, 17.
- (49) Takaya, T.; Su, C.; De La Harpe, K.; Crespo-Hernández, C. E.; Kohler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10285.
- (50) Olaso-González, G.; Merchán, M.; Serrano-Andrés, L. *J. Am. Chem. Soc.* **2009**, *131*, 4368.
- (51) Strickler, S. J.; Berg, R. A. *J. Chem. Phys.* **1962**, *37*, 814.
- (52) Rubio-Pons, O.; Serrano-Andrés, L.; Merchán, M. *J. Phys. Chem. A* **2001**, *105*, 9664.

- (53) Veryazov, V.; Widmark, P.-O.; Serrano-Andrés, L.; Lindh, R.; Roos, B. O. *Int. J. Quantum Chem.* **2004**, *100*, 626.
- (54) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Pedersen, T.; Pitonak, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224.
- (55) Noguera, M.; Blancafort, L.; Sodupe, M.; Bertran, J. *Mol. Phys.* **2006**, *104*, 925.
- (56) Fleig, T.; Knecht, S.; Hättig, C. *J. Phys. Chem. A* **2007**, *111*, 5482.
- (57) Zgierski, M. Z.; Patchkovskii, S.; Lim, E. C. *Can. J. Chem.* **2007**, *85*, 124.
- (58) Cohen, B. J.; Goodman, L. *J. Am. Chem. Soc.* **1965**, *87*, 5487.
- (59) Lavík, J.; Jelínek, O.; Plášek, J. *Photochem. Photobiol.* **1979**, *29*, 491.
- (60) Sobolewski, A. L.; Domcke, W. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2763.
- (61) Fülischer, M. P.; Serrano-Andrés, L.; Roos, B. O. *J. Am. Chem. Soc.* **1997**, *119*, 6168.
- (62) Lower, S. K.; El-Sayed, M. A. *Chem. Rev.* **1966**, *66*, 199.
- (63) Ullrich, S.; Schultz, T.; Zgierski, M. Z.; Stolow, A. *J. Am. Chem. Soc.* **2004**, *126*, 2262.
- (64) Ullrich, S.; Schultz, T.; Zgierski, M. Z.; Stolow, A. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2796.
- (65) Climent, T.; González-Luque, R.; Merchán, M.; Serrano-Andrés, L. *J. Phys. Chem. A* **2006**, *110*, 13584.
- (66) Serrano-Pérez, J. J.; Merchán, M.; Serrano-Andrés, L. *Chem. Phys. Lett.* **2007**, *434*, 107.
- (67) Gustavsson, T.; Bányász, A.; Lazzarotto, E.; Markovitsi, D.; Scalamani, G.; Frisch, M. J.; Barone, V.; Improta, R. *J. Am. Chem. Soc.* **2006**, *128*, 607.
- (68) Hare, P. M.; Crespo-Hernández, C. E.; Kohler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 435.
- (69) Barenboim, G. M.; Domanskii, A. N. *Luminescence of biopolymers and cells*; Plenum Press: New York, 1969; p 85.
- (70) Snipes, W. *Electron spin resonance and the effects of radiation on biological systems*; National Academy of Sciences: Washington, DC, 1966, p55.
- (71) Gupron, M.; Shulman, R. G.; Eisinger, J. *Proc. Natl. Acad. Sci. U.S.A.* **1966**, *56*, 814.
- (72) Marian, C. M. *J. Phys. Chem. A* **2007**, *111*, 1545.
- (73) Nikogosyan, D. N.; Letokhov, L. S. *Riv. Nuov. Cim.* **1983**, *6*, 1.
- (74) Nguyen, M. T.; Zhang, R.; Nam, P.-C.; Ceulemans, A. *J. Phys. Chem. A* **2004**, *108*, 6554.
- (75) Blancafort, L. *Photochem. Photobiol.* **2007**, *83*, 603.
- (76) Merchán, M.; Serrano-Andrés, L. *J. Am. Chem. Soc.* **2003**, *125*, 8108.
- (77) Kudrya, V. Y.; Yashchuk, V. M.; Levchenko, S. M.; Melnik, V. I.; Zaika, L. A.; Govorun, D. M. *Mol. Cryst. Liq. Cryst.* **2008**, *497*, 425.

CT100164M

JCTC

Journal of Chemical Theory and Computation

Harmonic and Anharmonic Vibrational Frequency Calculations with the Double-Hybrid B2PLYP Method: Analytic Second Derivatives and Benchmark Studies

Malgorzata Biczysko,^{*,†,‡} Pawel Panek,^{‡,§} Giovanni Scalmani,^{||} Julien Bloino,^{†,‡} and Vincenzo Barone^{*,†,⊥}

Scuola Normale Superiore, piazza dei Cavalieri 7, 56126 Pisa, Italy, Dipartimento di Chimica "Paolo Corradini" and CR-INSTM Village Università di Napoli Federico II, Complesso Univ. Monte S. Angelo, via Cintia, 80126 Napoli, Italy, Faculty of Chemistry, University of Wrocław, ul. Joliot-Curie 14, 50-383 Wrocław, Poland, Gaussian, Inc., 340 Quinnipiac St., Bldg 40, Wallingford, Connecticut 06492, and Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Pisa, Pisa, Italy

Received April 20, 2010

Abstract: This work aims to provide reliable benchmark data on the accuracy of harmonic and anharmonic vibrational frequencies computed with the B2PLYP double-hybrid density functional method. The exchange-correlation contributions required for the B2PLYP analytical second derivatives are presented here, which allow for the effective calculation of harmonic frequency as well as cubic and semidiagonal quartic force fields. The latter, in turn, are necessary to compute the anharmonic vibrational frequencies with the perturbative approach (VPT2). The quality of harmonic vibrational frequencies computed in conjunction with basis sets of double- to quadruple- ζ quality has been checked against reference data from the F38 benchmark set. Then, for an additional set of small closed- and open-shell systems, both harmonic frequencies and anharmonic contributions computed at the B2PLYP/N07D and the B2PLYP/aug-cc-pVTZ levels have been compared to their CCSD(T) counterparts. Moreover, for selected medium-size molecules (furan, pyrrole, thiophene, uracil, anisole, phenol, and pyridine), anharmonic frequencies have been compared to well established experimental results. Such benchmark studies have shown that the B2PLYP/N07D model provides good quality harmonic frequencies and describes correctly anharmonic contributions, the latter being of similar accuracy to their B3LYP/N07D counterparts, but obtained at significantly larger computational cost. Additionally, increased accuracy can be obtained by adopting hybrid models where the B2PLYP/N07D anharmonic contributions are combined with harmonic frequencies computed with more accurate quantum mechanical (QM) approaches or by B2PLYP with larger basis sets. This work confirmed also that most of the recently developed density functionals are significantly less suited for vibrational computations, while the B2PLYP method can be recommended for spectroscopic studies where a good accuracy of vibrational properties is required.

1. Introduction

Computational chemistry experiments have already been proven to deliver highly accurate results for small mole-

cules,^{1–5} clearly demonstrating their usefulness as tools for the prediction and understanding of many kinds of spectroscopic properties of molecular systems. At present, it is widely recognized that, for semirigid molecules, the computation of vibrational frequencies by a second-order perturbative approach (VPT2)^{6,7} can be applied even for quite large systems to support reliable interpretation of spectroscopic measurements. In particular, VPT2 computations coupled with semidiagonal quartic force fields evaluated at the CCSD(T) (coupled clusters with single, double, and

* To whom correspondence should be addressed. E-mail: malgorzata.biczysko@sns.it, vincenzo.barone@sns.it.

† Scuola Normale Superiore.

‡ Complesso Univ. Monte S. Angelo.

§ University of Wrocław.

|| Gaussian, Inc.

⊥ INFN.

perturbative inclusion of triple excitations⁸) level in conjunction with basis sets of at least triple- ζ quality usually provide results with an accuracy on the order of 10–15 cm^{-1} for the fundamental transitions.^{9–21} However, computations at the CCSD(T) level are still limited to small systems, so that the extension of accurate computational studies to larger systems requires cheaper yet reliable electronic structure methods. In this respect, the density functional theory (DFT) stands as a valuable route, and several VPT2 computations based on the DFT anharmonic force fields have been reported for small and medium-sized semirigid molecules.^{22–27} Among the functionals tested, hybrid ones provide satisfactory results when coupled to basis sets of at least double- ζ plus polarization quality supplemented by diffuse sp functions. However, as we recently pointed out,^{28,29} computation of the vibrational frequencies turned out to be a particularly challenging task, even for newly developed density functionals. As a matter of fact, some of the most successful last-generation functionals (M06-2X and ω B97X) provided quite disappointing results, showing that vibrational properties should not be overlooked while optimizing parameters in this kind of functional.

Recently, some of us have presented a DFT/N07D model^{30–32} which, for density functionals like B3LYP,³³ CAM-B3LYP,³⁴ and PBE0,³⁵ provides results of remarkable quality for a broad range of spectroscopic parameters (ESR, IR, UV, ECD).^{36–38} In the search for a computational approach able to reproduce different spectroscopic properties with consistent accuracy, the double-hybrid B2PLYP³⁹ method appears as a promising alternative, as it has already been shown to provide accurate results even for excited electronic states,⁴⁰ including challenging topics like electronic circular dichroism.⁴¹ In this work, we test the performance of B2PLYP in the evaluation of vibrational properties, which represents an issue for several functionals, preventing their systematic use in computational spectroscopy. For this purpose, both harmonic and anharmonic vibrational frequencies will be computed using the B2PLYP approach. At this point, it should be remembered that anharmonic VPT2 computations require cubic and semidiagonal quartic force fields, which in turn can be effectively determined via numerical differentiation of analytically evaluated force constants.^{7,13,42,43} In this respect, anharmonic computations with the B2PLYP method have become feasible thanks to the development and implementation of the B2PLYP analytical 2nd derivatives (see section 2). Concerning the validation of the B2PLYP vibrational properties, and further extension of the DFT/N07D model, VPT2 computations with the B2PLYP method have been performed in conjunction with the N07D basis set for all systems. It should be noted that, in the case of B3LYP computations, the basis set extension beyond N07D has a negligible effect on the accuracy of vibrational properties.²⁹ However, in the case of B2PLYP, it can be expected that significantly larger basis sets are required due to the MP2 contribution. For this purpose, the quality of the B2PLYP/N07D harmonic frequencies has been assessed by comparison with the results obtained at the CCSD(T) level and from experimental data, while the basis set convergence has been checked by

extending the basis set to aug-cc-pVTZ (AVTZ) and/or aug-cc-pVQZ (AVQZ). Next, the performance of the B2PLYP/N07D model in evaluating the anharmonic contributions has been tested for a set of small closed- and open-shell systems by comparison to quartic force fields at the B2PLYP/AVTZ and CCSD(T) levels of theory. Moreover, for larger systems, the quality of the anharmonic frequencies computed with the B2PLYP/N07D and hybrid models has been assessed relative to state-of-the-art experimental results. Finally, the accuracy of the vibrational properties computed with several other density functionals has been evaluated, in order to validate further the conclusions drawn on the basis of a smaller set of data.^{28,29}

The paper is organized as follows. Section 2 describes the exchange-correlation contributions required by the formalism of the analytical second derivatives of the B2PLYP energies. The details on the computational models applied to the determination of structures and harmonic and anharmonic vibrations are gathered in Section 3. Section 4 reports the benchmark results on vibrational properties computed at the B2PLYP/N07D level. Harmonic frequencies for the molecules from the F38 benchmark set are reported in section 4.1. VPT2 computations for small closed- and open-shell systems validated by comparison with accurate results at the CCSD(T) level are collected in section 4.2. Additionally, B2PLYP/N07D and hybrid B2PLYP//AVTZ/N07D VPT2 anharmonic frequency results are compared to experimental data in section 4.3. Finally, our conclusions on the accuracy of B2PLYP and other DFT approaches in computing vibrational properties are presented in section 4.4.

2. Exchange-Correlation Contributions to the B2PLYP Analytic Second Derivative

B2PLYP belongs to the family of so-called “double-hybrid” methods, which are essentially a second-order perturbation (PT2) treatment of the correlation energy. When the results of a Hartree–Fock (HF) self-consistent field (SCF) calculation are used as a zeroth-order reference, the PT2 approach corresponds to the well-known MP2 method. However, the results of a DFT Kohn–Sham (KS) SCF can be used as a reference as well, and by a suitable (semiempirical) scaling of the PT2 contribution to the energy, a significant improvement in the accuracy of the method can be achieved.^{39,44} Therefore, the formalism of the derivatives of the KS-PT2 method (and of B2PLYP in particular) is best understood as a combination of the KS-SCF and MP2 first and second derivatives.

In the following, we will describe only and all of the exchange-correlation (XC) terms required to evaluate the second derivatives of the KS-PT2 energy. The interested reader is invited to review refs 45–47, which describe the details of the evaluation of the first and second derivatives of the KS-SCF energy, and refs 48–50, where the MP2 first derivatives are illustrated. Moreover, a presentation of the overall formalism of the MP2 energy second derivatives can be found in refs 51–53, while the first derivatives of the KS-PT2 method have been recently reported in ref 54. Finally, the exchange-correlation terms involved in any “post-KS” gradient (and thus also in the B2PLYP gradient) have been

described in detail in connection with the implementation of the time-dependent DFT (TD-DFT) gradient in ref 55. The complete formalism of the KS-PT2 2nd derivatives, including frozen-core approximation and solvent effects by means of the polarizable continuum model (PCM),^{56,57} will be presented in a more organic form in a forthcoming paper.

In order to concisely write the various exchange-correlation (XC) contributions to the KS-PT2 energy and its derivatives, it is necessary to introduce some notation. First, we write the XC contribution to the KS-SCF energy as

$$E_{xc} = w \mathcal{F}[\mathbf{v}_I] \quad (1)$$

where we assume an integration grid point index on both the integration weights w and the functional values \mathcal{F} , and we assume the required sum over the grid points. The functional itself depends on a set of variables $\{\mathbf{v}_I\}$ ⁵⁵ which typically include the density ρ , the density gradient $\nabla\rho$, the kinetic energy density τ , and the Laplacian of the density $\nabla^2\rho$, i.e.,

$$\{\mathbf{v}_I\} = \{\rho, \nabla\rho, \tau, \nabla^2\rho\} \quad (2)$$

Note that all of the elements of the set $\{\mathbf{v}_I\}$ are linear in a one-particle density matrix \mathbf{P} according to

$$\mathbf{v}_I[\mathbf{P}] = \sum_{\mu\nu} P_{\mu\nu} \mathbf{v}_{I,\mu\nu} \quad (3)$$

where $\mathbf{v}_{1,\mu\nu} = \chi_\mu \chi_\nu$ for the density, $\mathbf{v}_{1,\mu\nu} = \nabla(\chi_\mu \chi_\nu)$ for the density gradients, $\mathbf{v}_{1,\mu\nu} = (\nabla\chi_\mu) \cdot (\nabla\chi_\nu)$ for the kinetic energy density, and $\mathbf{v}_{1,\mu\nu} = \nabla^2(\chi_\mu \chi_\nu)$ for the Laplacian of the density. The set $\{\chi_\mu\}$ represents the atomic orbital (AO) basis set. The functional \mathcal{F} is usually written as depending on the squared norms of the density gradient, i.e., $\gamma_{\sigma\sigma'} = (\nabla\rho_\sigma) \cdot (\nabla\rho_{\sigma'})$, where σ and σ' are spin labels. However, for the sake of a more concise notation, we will assume that the chain rule has been applied to the functional derivatives to obtain derivatives with respect to the elements of the set $\{\mathbf{v}_I\}$.

The first derivative of the XC energy is well-known⁴⁵ to be

$$E_{xc}^x = w^x \mathcal{F} + w \mathcal{F}^I \mathbf{v}_I^{(x)} \quad (4)$$

where w^x represents the first derivatives of the integration weights, \mathcal{F}^I is the first derivative of the functional with respect to the I th variable, and a sum over I is implied. Also, with the parentheses, we indicate the *explicit* dependence of the variables through the basis function, i.e.,

$$\mathbf{v}_I^{(x)} = \sum_{\mu\nu} P_{\mu\nu} \mathbf{v}_{I,\mu\nu}^x \quad (5)$$

The corresponding XC energy second derivatives^{45,46} can be written as

$$E_{xc}^{xy} = w^{xy} \mathcal{F} + w^x \mathcal{F}^I \mathbf{v}_I^{(y)} + w^y \mathcal{F}^I \mathbf{v}_I^{(x)} + w \mathcal{F}^{IJ} \mathbf{v}_I^{(x)} \mathbf{v}_J^{(y)} + w \mathcal{F}^I \mathbf{v}_I^{(x,y)} + w^x \mathcal{F}^I \mathbf{v}_I^{[y]} + w \mathcal{F}^{IJ} \mathbf{v}_I^{(x)} \mathbf{v}_J^{[y]} + w \mathcal{F}^I \mathbf{v}_I^{(x)[y]} \quad (6)$$

where the first five terms on the right-hand side involve *explicit* dependence of the weights and the variables on the perturbations, while the last three terms account for the

implicit dependence of the variables on the density derivative, which we indicate using the square brackets

$$\mathbf{v}_I^{[x]} = \mathbf{v}_I[\mathbf{P}^x] = \sum_{\mu\nu} P_{\mu\nu}^x \mathbf{v}_{I,\mu\nu} \quad (7)$$

Note that the last three terms in eq 6 can be also written as follows

$$\sum_{\mu\nu} P_{\mu\nu}^y (w^x \mathcal{F}^I \mathbf{v}_{I,\mu\nu} + w \mathcal{F}^{IJ} \mathbf{v}_J^{(x)} \mathbf{v}_{I,\mu\nu} + w \mathcal{F}^I \mathbf{v}_{I,\mu\nu}^x) = \langle \mathbf{P}^y \mathbf{G}_{xc}^{(x)} \rangle \quad (8)$$

i.e., like the trace of the density derivative \mathbf{P}^y with the XC portion of the skeleton Fock matrix derivative. The occupied-virtual block of the density derivative \mathbf{P}_{ov}^x is the solution of the couple-perturbed KS (CP-KS) equations, whose right-hand side involves the skeleton Fock matrix derivative and the additional XC term

$$\mathbf{G}_{xc}[\mathbf{P}_{oo}^x] = -w \mathcal{F}^{IJ} \mathbf{v}_J[\mathbf{S}_{oo}^x] \mathbf{v}_{I,\mu\nu} \quad (9)$$

where⁵⁸ $\mathbf{S}_{oo}^x = -\mathbf{P}_{oo}^x$, while the left-hand side includes the corresponding XC term, which depends on the unknown quantity \mathbf{P}_{ov}^x , i.e.

$$\mathbf{G}_{xc}[\mathbf{P}_{ov}^x] = w \mathcal{F}^{IJ} \mathbf{v}_J[\mathbf{P}_{ov}^x] \mathbf{v}_{I,\mu\nu} \quad (10)$$

In addition to the terms in eq 6, there is also an XC contribution to the total energy second derivatives, namely, through the $\langle \mathbf{W}^y \mathbf{S}^x \rangle$ trace, where the derivatives of the energy-weighted matrix express the dependence of the SCF orbital energies on the perturbations, which is assembled from the complete \mathbf{P}^x and \mathbf{F}^x matrices.

On the other hand, the XC contributions to the KS-PT2 gradient,⁵⁴ or more generally speaking to any “post-KS” gradient,⁵⁵ assume the following form:

$$E_{xc}^{\text{KS-PT2}(x)} = w^x \mathcal{F} + w \mathcal{F}^I \mathbf{v}_I^{(x)} + \langle \boldsymbol{\gamma} \mathbf{G}_{xc}^{(x)} \rangle - \langle \mathbf{S}^y \mathbf{G}_{xc}[\boldsymbol{\gamma}] \rangle = w^x (\mathcal{F} + \mathcal{F}^I \mathbf{v}_I[\boldsymbol{\gamma}]) + w (\mathcal{F}^I + \mathcal{F}^{IJ} \mathbf{v}_J[\boldsymbol{\gamma}]) \mathbf{v}_I^{(x)} + w \mathcal{F}^I \mathbf{v}_I^{(x)}[\boldsymbol{\gamma}] - w \mathcal{F}^{IJ} \mathbf{v}_J[\boldsymbol{\gamma}] \mathbf{v}_I[\mathbf{S}^y] \quad (11)$$

where $\boldsymbol{\gamma}$ is the correlation contribution to the one-particle density matrix, back-transformed to the AO basis. The occupied-virtual block of $\boldsymbol{\gamma}$ is found by solving the so-called Z-vector equations. These are CP-KS equations whose right-hand side involves the KS-PT2 Lagrangian,⁵⁴ which is indeed identical to the MP2 Lagrangian⁴⁸ since the KS-PT2 energy $E^{\text{KS-PT2}}$ does not involve any explicit XC energy term beyond the KS-SCF level. The last term on the right-hand side of eq 11 represents the trace of the overlap matrix derivative \mathbf{S}^x with the correlation contribution to the appropriate energy-weighted density matrix $\mathbf{W}^{\text{KS-PT2}}$.

Finally, the XC contributions to the KS-PT2 energy second derivatives are

$$\begin{aligned}
E_{xc}^{\text{KS-PT2}(x,y)} = & w^{xy}(\mathcal{F} + \mathcal{F}^{\text{I}}\mathbf{v}_j[\gamma]) + w^x[(\mathcal{F}^{\text{I}} + \mathcal{F}^{\text{I,J}}\mathbf{v}_j[\gamma])\mathbf{v}_i^{(y)} + \\
& \mathcal{F}^{\text{I}}\mathbf{v}_i^{(y)}[\gamma]] + w^y[(\mathcal{F}^{\text{I}} + \mathcal{F}^{\text{I,J}}\mathbf{v}_j[\gamma])\mathbf{v}_i^{(x)} + \mathcal{F}^{\text{I}}\mathbf{v}_i^{(x)}[\gamma]] + \\
& w[(\mathcal{F}^{\text{I,J}} + \mathcal{F}^{\text{I,J,K}}\mathbf{v}_k[\gamma])\mathbf{v}_i^{(x)}\mathbf{v}_j^{(y)} + \mathcal{F}^{\text{I,J}}(\mathbf{v}_i^{(x)}\mathbf{v}_j^{(y)})[\gamma] + \\
& \mathbf{v}_i^{(y)}\mathbf{v}_j^{(x)}[\gamma]] + (\mathcal{F}^{\text{I}} + \mathcal{F}^{\text{I,J}}\mathbf{v}_j[\gamma])\mathbf{v}_i^{(x,y)} + \mathcal{F}^{\text{I}}\mathbf{v}_i^{(x,y)}[\gamma]] + \\
& \langle \mathbf{P}^y \mathbf{G}_{xc}^{(x)}[\gamma] \rangle + \langle \gamma^y \mathbf{G}_{xc}^{(x)} \rangle - \langle \mathbf{S}^{xy} \mathbf{G}_{xc}[\gamma] \rangle - \langle \mathbf{S}^x \mathbf{G}_{xc}^{(y)}[\gamma] \rangle - \\
& \langle \mathbf{S}^x \mathbf{G}_{xc}^{[y]}[\gamma] \rangle - \langle \mathbf{S}^x \mathbf{G}_{xc}[\gamma^y] \rangle \quad (12)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{G}_{xc}^{(x)}[\gamma] = & w^x(\mathcal{F}^{\text{I}} + \mathcal{F}^{\text{I,J}}\mathbf{v}_j[\gamma])\mathbf{v}_{i,\mu\nu} \\
& + w[(\mathcal{F}^{\text{I,J}} + \mathcal{F}^{\text{I,J,K}}\mathbf{v}_k[\gamma])\mathbf{v}_j^{(x)} + \mathcal{F}^{\text{I,J}}\mathbf{v}_j^{(x)}[\gamma]]\mathbf{v}_{i,\mu\nu} \\
& + w(\mathcal{F}^{\text{I}} + \mathcal{F}^{\text{I,J}}\mathbf{v}_j[\gamma])\mathbf{v}_{i,\mu\nu}^x \quad (13)
\end{aligned}$$

and

$$\mathbf{G}_{xc}^{[x]}[\gamma] = w\mathcal{F}^{\text{I,J,K}}\mathbf{v}_k[\gamma]\mathbf{v}_j[\mathbf{P}^x]\mathbf{v}_{i,\mu\nu} \quad (14)$$

In order to completely evaluate eq 12, the full derivative γ^x of the correlation contribution to the one-particle density matrix must be computed. The occupied–occupied and virtual–virtual blocks of γ^x depend on products of PT2 amplitudes and amplitudes derivatives. The latter can be assembled from undifferentiated amplitudes and orbital energies, two-electron integral derivatives, and the derivatives of the Fock operators (see, e.g., eq 39 in ref 53). These are the derivatives of the Fock matrix in the canonical molecular orbital basis, which are no longer diagonal matrices and include automatically the proper XC contributions, once the nuclear coordinate CP-KS equations have been solved and the full \mathbf{P}^x and \mathbf{F}^x matrices are available. Thus, the only remaining piece is the occupied-virtual block of γ^x , which is the solution of the *derivative* Z-vector equations, whose right-hand side involves the derivatives of the MP2 Lagrangian^{51–53} and all the terms from the derivatives of the left-hand side which do not involve the unknowns γ_{ov}^x , i.e., the quantities in eqs 13 and 14 together with the additional term

$$w\mathcal{F}^{\text{I,J}}\mathbf{v}_j[\gamma_{oo}^x + \gamma_{vv}^x]\mathbf{v}_{i,\mu\nu} \quad (15)$$

3. Computational Details

Density functional theory computations have been carried out using the double-hybrid B2PLYP³⁹ method in conjunction with the recently developed polarized double- ζ N07D^{30–32,59} and aug-cc-pVXZ (X = T, Q)^{60,61} basis sets. The N07D basis set has been constructed by adding a reduced number of polarization and diffuse functions to the 6-31G set (see refs 30 and 31 for details), leading to an optimum compromise between reliability and computational cost.

All structures have been optimized using tight convergence criteria, followed by the computation of the anharmonic frequencies by means of the VPT2 approach,^{6,7} as implemented in the Gaussian package.⁶² Semidiagonal quartic force fields have been evaluated by numerical differentiation (with a standard 0.025 Å step) of analytical second derivatives.⁴² Since VPT2 computations are sensitive to the proper treatment of the Fermi resonances, it is crucial to automatically neglect nearly singular contributions (deperturbed

computations). This is performed by effectively removing interactions in the second-order treatment, which are more properly treated in the first-order. For this purpose, our VPT2 implementation⁷ makes use of the criteria proposed by Martin and Boese,²⁵ through an automated scheme that has already been shown to provide accurate results, at least for fundamental bands.⁶³ Additionally, in some cases, the hybrid CCSD(T)/DFT or DFT AVTZ/N07D approaches have also been applied to evaluate the anharmonic frequencies, and two possible routes have been implemented. In the simpler one (DPT2), the harmonic frequencies computed at the higher level of theory (CCSD(T), B2PLYP/AVTZ) are *a posteriori* corrected by the anharmonic contributions ($\Delta\nu$) derived from VPT2 computations performed at the lower level: $\nu_{\text{Higher/Lower}} = \omega_{\text{Higher}} + \Delta\nu_{\text{Lower}}$. Such an approximation, in particular within the CCSD(T)/DFT scheme, has been already validated for several closed- and open-shell systems (see for instance refs 29, 64–67). The second route introduces the harmonic frequencies evaluated at the higher level directly into the VPT2 computations along with the 3rd and 4th order force constants obtained at the lower level of theory. Such an approach is available in the Gaussian package through the InDerAU and InFreq options, with harmonic frequencies computed at the higher level of theory listed in the input stream (a feature available in the standard package⁶⁸) or with the corresponding Hessian matrix read from the checkpoint file. For the latter case, an automatic procedure which compares normal modes computed by the two levels of theory and replaces harmonic data accordingly is introduced in this work. Such an implementation facilitates the application of a hybrid InFreq route for large systems for which the ordering of several closely lying vibrations might be exchanged. It should be noted that the InFreq procedure might significantly improve the quality of the results in difficult cases, i.e., when large discrepancies between harmonic frequencies computed at two levels of theory or Fermi resonances are present.

In addition to the computations with the B2PLYP method, we decided to benchmark the performances of other density functionals, in order to confirm the findings obtained in several case studies^{28,29} where an unsatisfactory description of vibrational frequencies had been found out. In this context, a broad range of recently introduced density functionals, namely, M06/M06-2X,^{69,70} the ω B97 family,^{71,72} HSE06,⁷³ and LC- ω PBE,⁷⁴ has been considered. For the sake of completeness, standard functionals like B3LYP,³³ CAM-B3LYP,³⁴ and B97-1⁷⁵ and the parameter-free PBE0³⁵ have also been included in our tests. All calculations have been performed with a locally modified version of the Gaussian suite of quantum chemistry programs.⁶²

4. Validation of the B2PLYP Method for the Calculation of Vibrational Frequencies

4.1. Harmonic Vibrational Frequencies for Small Molecules from the F38 Database. The present work is devoted to the validation of the B2PLYP/N07D model for the computation of vibrational frequencies. Thus, it is appropriate to start the analysis discussing the accuracy of

Table 1. Harmonic (ω) Vibrational Frequencies (in cm^{-1}) Computed with the B2PLYP and B3LYP Functionals and the N07D or aug-cc-pVTZ(AVTZ) (for B2PLYP also aug-cc-pVQZ(AVQZ)) Basis Sets for Molecules from the F38 Benchmark Set, and Compared to the F38 Reference Data

		B2PLYP			B3LYP		
		exp. ^a	N07D	AVTZ	AVQZ	N07D	AVTZ
H ₂	ω_1	4401	4501	4464	4461	4451	4418
CH ₄	ω_1	1367	1361	1353	1352	1342	1339
	ω_2	1583	1576	1576	1575	1557	1557
	ω_3	3026	3068	3050	3050	3037	3028
	ω_4	3157	3191	3162	3163	3150	3130
NH ₃	ω_1	3478	3517	3489	3492	3485	3469
	ω_2	1084	1028	1037	1034	999	1025
	ω_3	3597	3660	3617	3621	3621	3588
	ω_4	1684	1682	1673	1674	1666	1664
H ₂ O	ω_1	1649	1652	1635	1637	1641	1627
	ω_2	3832	3832	3813	3823	3814	3796
	ω_3	3943	3951	3924	3934	3922	3899
HF	ω_1	4139	4096	4099	4107	4071	4070
CO	ω_1	2170	2155	2154	2161	2205	2207
N ₂	ω_1	2359	2351	2341	2346	2453	2448
F ₂	ω_1	917	970	1016	1012	1023	1050
C ₂ H ₂	ω_1	624	588	643	649	622	666
	ω_2	747	765	766	762	772	770
	ω_3	2008	2025	2024	2025	2063	2068
	ω_4	3415	3457	3429	3432	3429	3412
	ω_5	3495	3550	3530	3524	3531	3517
HCN	ω_1	727	753	745	745	768	759
	ω_2	2127	2129	2125	2129	2198	2200
H ₂ CO	ω_3	3443	3495	3460	3456	3473	3444
	ω_1	2937	2951	2930	2928	2901	2885
	ω_2	1778	1790	1782	1786	1819	1813
	ω_3	1544	1543	1538	1540	1529	1530
	ω_4	1188	1192	1201	1204	1188	1198
CO ₂	ω_5	3012	3023	2992	2991	2967	2940
	ω_6	1269	1269	1268	1272	1260	1263
	ω_1	673	660	666	668	666	674
	ω_2	1353	1343	1341	1345	1370	1369
	ω_3	2392	2400	2384	2392	2416	2400
N ₂ O	ω_1	596	572	599	608	592	617
	ω_2	1298	1310	1298	1301	1337	1324
	ω_3	2282	2271	2259	2279	2352	2340
Cl ₂	ω_1	560	540	551	555	532	537
OH	ω_1	3738	3758	3737	3748	3712	3695
	MIN		-56	-47	-50	-85	-72
	MAX		100	99	95	106	133
	MUE		23	18	17	33	33

^a Benchmark harmonic frequency values as compiled in refs 70 and 76 on the basis of data from refs 78–80.

harmonic frequencies with reference to the recently introduced benchmark set F38,⁷⁰ designed to cover a broad range of frequencies for small molecules. It has been applied here to assess the accuracy of harmonic vibrational frequencies for several density functional^{70,76,77} methods. The F38 reference set of data is based on the best experimental harmonic frequencies,^{78,79} with the single exception for the umbrella mode of the NH₃, which is taken from a CCSD(T)/cc-pVQZ calculation.⁸⁰ It should be noted that, for consistency with the available benchmark studies, Table 1 compares harmonic frequencies computed at the B2PLYP level to the original F38 reference data,⁷⁰ while in section 4.2 we report the best theoretical harmonic frequencies up to date for some molecules from the F38 database. However, the best experimental and theoretical values for these molecules (H₂O,⁸¹ NH₃,⁸³ and H₂CO⁸³) are very similar, with an

average deviation of 6 cm^{-1} only. The results presented in Table 1 clearly show the good overall accuracy of the harmonic frequencies computed by the B2PLYP/N07D model, which are off by only 1.5% on average from the reference, with a maximum error of about 5%. In absolute values, this corresponds to a mean unsigned error (MUE) of about 23 cm^{-1} , and maximum negative (MIN) and positive (MAX) discrepancies of -56 cm^{-1} and 100 cm^{-1} , respectively, with the single absolute deviation above 60 cm^{-1} observed for the H–H stretching frequency in H₂. Additionally, it can be noted that slightly higher absolute deviations are observed for frequencies above 2500 cm^{-1} . The separate analysis performed for frequencies above and below this threshold led to MUEs of 37 cm^{-1} and 15 cm^{-1} , respectively. The results presented in Table 1 show also that the extension of the basis set up to aug-cc-pVTZ, or even aug-cc-pVQZ (with the exception of the F₂ molecule), leads in most cases to a slightly superior agreement with the reference data with a MUE of 18 cm^{-1} and 17 cm^{-1} , respectively. Such an effect is most pronounced for the frequencies above 2500 cm^{-1} , and in the extreme case of the H₂ molecule, the extension of the basis set improves the agreement by about 50%. It is worth adding that the frequencies computed with the aug-cc-pVTZ and aug-cc-pVQZ basis sets agree on average to 5 cm^{-1} , with a maximum discrepancy of 20 cm^{-1} , confirming that frequency calculations approach the basis set convergence at the AVTZ level. Indeed, the MP2 contribution to B2PLYP causes the computed harmonic frequencies to be not fully converged with respect to the basis set at the N07D level. However, comparison of the results obtained with the double- ζ N07D (58 basis functions for Cl₂) and aug-cc-pVQZ (168 basis functions for Cl₂) basis sets shows that the error compensation allows the B2PLYP/N07D model to deliver good quality harmonic frequencies. The above arguments are confirmed by the data gathered in Table 1, which point out that, for the standard B3LYP functional, no overall improvement is obtained going from N07D to the more computationally demanding AVTZ basis set, as already shown by the comparison of the harmonic frequencies computed with B3LYP using basis sets of both double- and triple- ζ quality.⁶⁵ Additionally, it can be observed that the B2PLYP method outperforms the B3LYP functional, in line with preliminary studies by Grimme.³⁹ In fact, for the F38 database, B2PLYP/N07D shows a MUE about 30% smaller than B3LYP/N07D (34 cm^{-1}). Thus, despite the fact that particularly difficult cases and/or a need of extreme accuracy might require CCSD(T) computations with extended basis sets, the overall impression is that the B2PLYP stands as the most accurate DFT model to compute harmonic frequencies.

4.2. Anharmonic Vibrational Frequencies for Small Closed- and Open-Shell Systems: B2PLYP vs CCSD(T). In a next step, we compare results provided by the B2PLYP method with those obtained at the CCSD(T) level, with extended basis sets, in order to dissect the overall accuracy of the vibrational frequencies into harmonic and anharmonic contributions. In this respect, we have chosen a set of closed- and open-shell molecules, for which the accuracy of CCSD(T) results has been confirmed by a comparison with experimental data.^{5,29,81–83,87–89} As we did in the previous

Table 2. Harmonic (ω) and Anharmonic (ν) Vibrational Frequencies (in cm^{-1}) Computed at the B2PLYP/N07D, B2PLYP/AVTZ, and Hybrid CC+DFT Levels for Selected Closed- and Open-Shell Systems, Compared to the Best Available Theoretical Results Computed at Coupled Cluster Levels

	B2PLYP			CCSD(T)				
	ω	ν	$\nu\text{CC+DFT}^a$	ω	ν	$\nu\text{CC+DFT}^a$	ω	ν
H ₂ O		N07D			AVTZ		CBS(67)/PES ^b	
ν_1	3832	3659	3663	3812	3645	3669	3836	3659
ν_2	1652	1598	1596	1635	1582	1598	1650	1596
ν_3	3951	3766	3761	3924	3744	3766	3946	3758
HCO		N07D			AVTZ		CBS/aCV ^c	CBS+ QZ ^c
ν_1	2724	2483	2476	2708	2458	2466	2717	2460
ν_2	1892	1868	1880	1886	1862	1881	1905	1878
ν_3	1120	1087	1088	1112	1077	1084	1120	1093
FCO		N07D			AVTZ		augVQZ ^d	augVTZ ^d
ν_1	1896	1834	1838	1888	1849	1861	1900	1864
ν_2	1019	978	1012	1037	1009	1026	1054	1025
ν_3	619	608	623	628	617	623	634	624
H ₂ CO		N07D			AVTZ			AVTZ(F12a) ^e
ν_1	2951	2794	2775	2930	2775	2778	2933	2784
ν_2	1790	1760	1747	1782	1752	1747	1777	1747
ν_3	1543	1509	1498	1538	1505	1499	1532	1498
ν_4	1192	1175	1170	1201	1184	1170	1187	1167
ν_5	3023	2842	2823	2992	2842	2853	3004	2849
ν_6	1268	1248	1247	1268	1246	1246	1267	1246
H ₂ O ₂		N07D			AVTZ			AVTZ(F12a) ^e
ν_1	3788	3598	3606	3777	3590	3609	3796	3606
ν_2	1431	1390	1395	1431	1386	1391	1436	1393
ν_3	909	877	881	930	901	883	913	880
ν_4	389	357	351	376	310	317	384	378
ν_5	3788	3601	3609	3777	3594	3613	3796	3608
ν_6	1324	1272	1278	1321	1262	1270	1329	1280
NH ₃		N07D			AVTZ			cc-pwCVQZ ^f
ν_1	3517	3372	3344	3488	3348	3348	3489	3342
ν_2	1028	938	986	1037	954	993	1076	1001
ν_3	3660	3490	3449	3617	3450	3452	3619	3444
ν_4	1682	1635	1633	1673	1626	1633	1680	1635
PH ₃		N07D			AVTZ			cc-pwCVQZ ^f
ν_1	2427	2328	2329	2427	2328	2329	2429	2331
ν_2	1031	1009	996	1018	998	997	1017	997
ν_3	2439	2329	2327	2437	2327	2327	2437	2336
ν_4	1152	1127	1122	1150	1123	1121	1147	1122
F ₂ CN		N07D			AVTZ		augVQZ ^g	aVQZ+ augVTZ ^g
ν_1	1809	1787	1790	1796	1775	1790	1811	1781
ν_2	960	946	960	967	953	960	974	957
ν_3	544	538	547	547	542	547	552	546
ν_4	673	667	674	687	681	673	679	673
ν_5	1252	1219	1262	1261	1228	1262	1295	1262
ν_6	493	489	497	500	496	497	501	496
NH ₃ ⁺		N07D			AVTZ			VQZ ^h
ν_1	3395	3252	3231	3372	3234	3237	3375	3231
ν_2	873	928	921	864	922	923	865	910
ν_3	3590	3419	3388	3552	3393	3400	3559	3388
ν_4	1557	1523	1517	1548	1507	1510	1551	1507
PH ₃ ⁺		N07D			AVTZ			VQZ ^h
ν_1	2501	2406	2402	2505	2419	2410	2497	2400
ν_2	745	667	673	748	674	678	751	670
ν_3	2577	2482	2474	2584	2492	2476	2568	2469
ν_4	2577	2471	2463	2584	2494	2478	2568	2469
ν_5	1058	1036	1032	1056	1037	1035	1054	1029
ν_6	1059	1035	1030	1056	1037	1035	1054	1029
C ₂ H ₃		N07D			AVTZ			AVTZ(PES/S) ⁱ
ν_1	3290	3267	3242	3267	3129	3105	3242	3108
ν_2	3211	3053	3016	3178	3024	3021	3174	3016
ν_3	3109	2946	2907	3077	2917	2910	3070	2901
ν_4	1679	1659	1590	1667	1647	1590	1610	1576
ν_5	1412	1378	1356	1405	1370	1355	1390	1355
ν_6	1069	1019	1014	1061	1010	1013	1064	1015
ν_7	728	695	683	716	681	683	717	688
ν_8	932	920	895	935	921	893	907	892
ν_9	821	809	787	828	812	784	799	793
MIN	-48	-63	-27	-39	-68	-61		
MAX	69	83	17	57	71	14		
MUE	16	18	4	10	11	4		

^a Anharmonic corrections at the B2PLYP/N07D level. ^b Ref 81. ^c Ref 87. ^d Ref 29. ^e Ref 83. ^f Ref 82. ^g Ref 5. ^h Ref 88. ⁱ Analytic harmonic frequencies and anharmonic results from VCI calculations using five-mode potential coupling based on a full-dimensional PES computed at the RCCSD(T)/aug-cc-pVTZ level. For details on PES/S, see ref 89.

section, we start the analysis of vibrational data by discussing harmonic frequencies. It should be noted that discrepancies

in the former term can be reduced by applying hybrid CC/DFT schemes, which are also presented in Table 2. First,

considering the accuracy of harmonic frequencies, it is immediately apparent that the conclusions drawn in section 4.1 are, in general terms, confirmed. Namely, the values computed at the B2PLYP/N07D level agree well with the most accurate calculations, with a MUE of 16 cm^{-1} , and further improvement (MUE of 10 cm^{-1}) can be achieved by using the aug-cc-pVTZ basis set. Similarly, anharmonic frequencies computed at the B2PLYP/N07D level show a MUE of 18 cm^{-1} , while the extension of the basis set to aug-cc-pVTZ leads to a MUE of 11 cm^{-1} . Additionally, a significant improvement is achieved through hybrid approaches with harmonic frequencies computed at the CCSD(T) level, which lead to a MUE of 4 cm^{-1} with respect to the anharmonic data computed entirely at the CCSD(T) level. Such a finding confirms the remarkable accuracy of the anharmonic force fields computed with the B2PLYP method, showing also that an improved accuracy can be achieved by using harmonic frequencies of coupled cluster quality. It should be underlined that both hybrid models, which differ by the computational cost associated with the size of the basis set, provide equally accurate results. This demonstrates clearly that the better agreement for the vibrational frequencies computed with the AVTZ basis set should be attributed uniquely to the higher accuracy of the harmonic component. In summary, a direct comparison with accurate computations at the CCSD(T) level clearly shows that the B2PLYP/N07D model provides harmonic frequencies of good accuracy and leads to a description of the anharmonic contributions in agreement with more accurate QM methods. However, it should be noted that results of equivalent accuracy can be delivered by hybrid approaches with anharmonic force fields obtained using the less computationally demanding B3LYP/N07D method.²⁹

4.3. Anharmonic Vibrational Frequencies of Larger Molecules Computed with B2PLYP/N07D and Hybrid Schemes. In this section, the performances of the B2PLYP method will be checked against well established experimental data for medium-size molecules. In this respect, we have chosen a set of organic aromatic systems, namely, pyridine, furan, pyrrole, thiophene, uracil, phenol, and anisole, for which previous calculations of anharmonic frequencies using the B3LYP or the B97-1 density functionals resulted in a very good agreement with the experimental results.^{23–25,63,84–86} In this work, both the B2PLYP/N07D model and a hybrid scheme with harmonic frequencies refined through B2PLYP/AVTZ calculations have been tested. For the latter, corrections have been applied to all normal modes or only to normal modes above 2500 cm^{-1} , in line with the findings reported in section 4.1 which displayed a larger basis set dependence for higher harmonic frequencies. Table 3 reports the mean unsigned errors with respect to the experimental data, along with maximum (negative and positive) deviations for all molecules considered. Figure 1 shows differences between computed and experimental frequencies for all normal modes of pyridine, furan, pyrrole, thiophene, uracil, phenol, and anisole, which are listed in order of increasing wavenumber. First, it can be observed that the overall agreement of the B2PLYP/N07D anharmonic frequencies with the reference data is very good, i.e., in the range of $9–15\text{ cm}^{-1}$ for all the

Table 3. Mean Absolute Errors (MUE), Maximum Negative (MIN) and Positive (MAX) Deviations of Anharmonic Vibrational Frequencies (in cm^{-1}) Computed with the B2PLYP/N07D and Hybrid B2PLYP/(AVTZ/N07D) Models As Compared to the Experimental Data^a

	N07D			AVTZ/N07D			AVTZ>2500 ^b / N07D		
	MUE	MIN	MAX	MUE	MIN	MAX	MUE	MIN	MAX
pyridine	9	-22	40	17	-35	53	10	-35	40
furan	9	-11	29	9	-14	55	5	-11	7
pyrrole	10	-7	32	11	-20	43	6	-11	28
thiophene	12	-10	34	6	-8	29	7	-10	21
uracil	11	-41	31	8	-12	27	9	-41	31
phenol	12	-18	70	13	-26	65	11	-27	70
anisole	15	-10	48	15	-51	72	12	-51	48
average all	11	-17	41	11	-24	49	9	-27	35

^a Experimental data are taken from (and references therein): pyridine, ref 90; furan and pyrrole, ref 91; thiophene, ref 92; uracil, ref 93; phenol, ref 94; anisole, ref 86. ^b Hybrid scheme applied to normal modes with frequencies above 2500 cm^{-1} , see text for details.

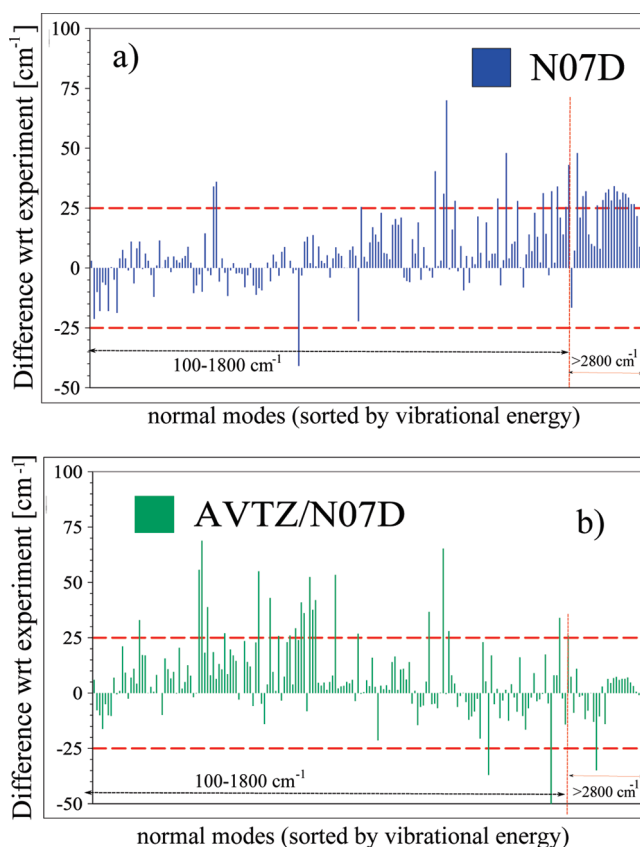


Figure 1. Performance of the B2PLYP/N07D (panel a) and hybrid B2PLYP/(AVTZ/N07D) (panel b) models for the prediction of anharmonic frequencies. The relative absolute discrepancies with respect to experimental results are shown for all normal modes of pyridine, furan, pyrrole, thiophene, uracil, phenol, and anisole, which are listed according to the increasing wavenumbers (in cm^{-1}).

molecules studied, corresponding to an average of 11 cm^{-1} . Moreover, as shown in Figure 1, despite maximum positive and negative discrepancies of 70 cm^{-1} and -41 cm^{-1} , respectively, the computed anharmonic frequencies are within 25 cm^{-1} of the experimental references, for almost all of the normal modes. Somewhat larger discrepancies are

Table 4. Mean Absolute Errors (MUE), Maximum Negative (MIN), and Positive (MAX) Deviations of Harmonic Vibrational Frequencies (in cm^{-1}) Computed with Several DFT/N07D Models for Molecules from the F38 Benchmark Set, and Compared to the F38 Reference Data

	MUE	MIN	MAX
CAM-B3LYP	52	-110	129
PBE0	50	-86	140
LC- ω PBE	74	-135	192
M06	57	-124	134
M06-2X	66	-77	163
HSE06	50	-86	138
ω B97	62	-93	159
ω B97X	60	-97	152
B97-1	33	-81	61

observed in the high frequency region of the spectrum. However, the relative deviation from experiment remains within 2% even for these frequencies, with only four frequencies above 1000 cm^{-1} exceeding this limit. It can be noted that a significant improvement of the absolute values in the high frequency region is achieved through a hybrid scheme, where the harmonic component is corrected using B2PLYP/aug-cc-pVTZ results. However, the hybrid scheme does not provide a systematic improvement for every normal mode; thus, the overall accuracy of the B2PLYP/(AVTZ/N07D) (referred to as AVTZ/N07D later on) model does not change with respect to the straightforward B2PLYP/N07D approach. On the other hand, it is possible to apply harmonic frequency refinements only to the high frequency normal modes ($>2500\text{ cm}^{-1}$); such a scheme effectively improves the agreement with respect to the experiment and should be considered when a good accuracy in the high frequency region of the spectrum is of particular importance. Thus, it can be concluded that the B2PLYP/N07D model provides very reliable anharmonic frequencies and can be safely applied to spectroscopic studies. However, it should also be

noted that the good overall accuracy of the B2PLYP/N07D and the hybrid AVTZ/N07D models is comparable to that obtained by less expensive anharmonic B3LYP/N07D calculations.

4.4. Accuracy of Harmonic and Anharmonic Vibrational Frequencies Computed with Other DFT/N07D Models. For the sake of completeness, we have investigated the performances of other density functional approaches using the N07D basis set. The same scheme as applied in the previous sections is used here. First, we assessed the accuracy of the harmonic frequencies with respect to the results from the F38 database. Table 4 collects the results obtained by means of some last generation DFT functionals not considered in the work of Zhao and Truhlar,^{70,77} along with a few standard functionals, which are among the most popular ones. First, it should be noted that the B3LYP, PBE0, M06, and M06-2X functionals together with the N07D basis set yield harmonic frequencies of accuracy essentially equivalent to the one reported in refs 70 and 77. Additionally, among all the density functionals tested either here or in the work by Zhao and Truhlar,^{70,77} only B3LYP, B97-1, and B2PLYP yield harmonic frequencies with the accuracy required for spectroscopic studies, and the B2PLYP method shows clearly the best results. As a next step, it seemed interesting to check also the quality of the cubic and semidiagonal quartic force fields computed with the recently developed density functionals. In this respect, the accuracy of anharmonic contributions has been assessed by comparison with their CCSD(T) counterparts for a few selected molecules, namely, H_2O , NH_3 , PH_3 , and F_2CN . This set of molecules has been chosen in view of the superior accuracy of anharmonic frequencies obtained with the hybrid CCSD(T)/B2PLYP scheme. The quality of the anharmonic force fields has been checked by inspection of the relative discrepancies between $\Delta\nu_{\text{PT}2}$'s computed at the DFT and CCSD(T) levels, respectively. Figure 2 shows a plot of the differences in $\Delta\nu_{\text{PT}2}$

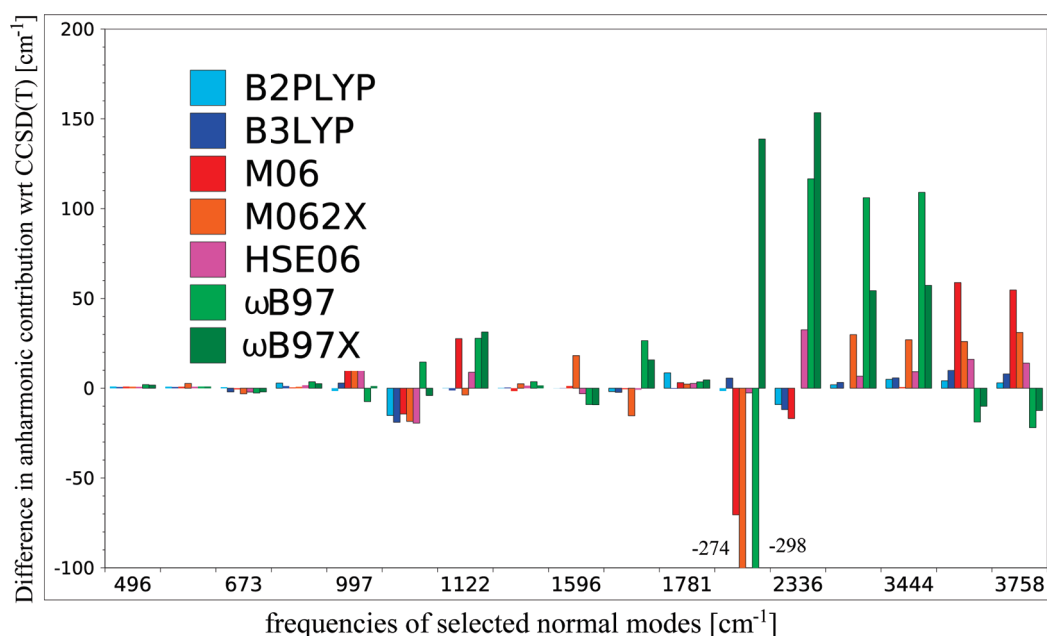


Figure 2. Performance of different density functionals for the prediction of the anharmonic contribution. The relative discrepancies with respect to the values computed at the CCSD(T) level are shown for each normal mode of H_2O , NH_3 , PH_3 , and F_2CN and are listed according to their wavenumbers (in cm^{-1}).

Table 5. Mean Absolute Errors (MUE), Maximum Negative (MIN) and Positive (MAX) Deviations of Anharmonic Vibrational Frequencies (in cm^{-1}) Computed with Several DFT/N07D Models and Compared to the Experimental Data^a

	pyridine	furan	pyrrole	thiophene	all
	MUE				
B2PLYP	9	9	10	12	10
B3LYP	9	5	6	7	7
CAM-B3LYP	19	22	21	20	20
PBE0	14	15	18	14	15
LC- ω PBE	28	44	47	35	39
M06	18	13	24	22	19
M06-2X	29	37	50	17	33
HSE06	12	16	20	11	15
ω B97	25	20	20	36	25
ω B97X	25	20	20	26	23
B97-1	13	5	5	9	8
	MIN				
B2PLYP	-22	-11	-7	-10	-13
B3LYP	-33	-15	-16	-19	-21
CAM-B3LYP	-9	6	7	3	2
PBE0	-19	-7	-4	-9	-10
LC- ω PBE	0	14	8	9	8
M06	-39	-17	-118	-63	-59
M06-2X	-18	-7	-15	-17	-14
HSE06	-20	-4	-4	-10	-9
ω B97	-53	-10	-21	4	-20
ω B97X	-50	1	-10	3	-14
B97-1	-47	-14	-13	-20	-24
	MAX				
B2PLYP	40	29	32	34	34
B3LYP	24	5	24	14	17
CAM-B3LYP	46	49	57	51	51
PBE0	75	31	43	40	47
LC- ω PBE	86	92	85	95	90
M06	74	50	55	29	52
M06-2X	136	137	298	56	157
HSE06	71	39	44	38	48
ω B97	74	75	72	82	76
ω B97X	59	70	68	65	66
B97-1	22	4	18	8	13

^a Experimental data from refs (and references therein): pyridine, ref 90; furan and pyrrole, ref 91; thiophene, ref 92.

between DFT and CCSD(T) for each normal mode of selected molecules, which are listed according to their wavenumbers (in cm^{-1}). First, it is clear that anharmonic corrections at the B2PLYP level agree very well with the reference data, as discussed in section 4.2. Similar results can be observed for B3LYP, further supporting the well-known good quality of the B3LYP/N07D force fields. The other density functionals show different trends, considering that only HSE06 performs in a qualitatively correct way, while functionals belonging to the M06 and the ω B97 families provide unreliable anharmonic corrections. Finally, we assessed the overall accuracy of the anharmonic vibrational frequencies computed by all of the DFT/N07D models considered in this work. For this purpose, Table 5 reports the mean unsigned errors and maximum deviations with respect to experimental data for pyridine, furan, pyrrole, and thiophene. These results show clearly that, among last generation DFT models, only the B2PLYP method (as discussed above) provides anharmonic frequencies in good agreement with experimental results, consistent with the accuracy of harmonic contributions and anharmonic correc-

tions discussed above. Moreover, the good performances of the B3LYP and the B97-1 functionals, when used in conjunction with the N07D basis set, are confirmed. In fact, for both functionals, the MUE is lower than 8 cm^{-1} . Qualitatively correct frequencies are also predicted by the PBE0 and HSE06 functionals, both showing MUEs of about 15 cm^{-1} . All of the other DFT models considered yield MUEs in the range of $20\text{--}40 \text{ cm}^{-1}$ and also show larger absolute discrepancies. Overall, the results presented in this section show that most of the recently developed density functionals are significantly less accurate in the calculation of vibrational frequencies, confirming the conclusions drawn in refs 28 and 29, on the basis of a smaller benchmark set. On the other hand, the B2PLYP method should be preferred for spectroscopic studies where a good accuracy of the vibrational properties is required.

5. Conclusions

In this work, we presented a concise exposition of the formalism of the analytic second derivatives for the double-hybrid B2PLYP method, along with an assessment of their accuracy in the calculation of vibrational properties. To that end, the computed harmonic vibrational frequencies have been compared with the best experimental estimates from the established F38 benchmark set. Additionally, for several small closed- and open-shell systems, both harmonic frequencies and anharmonic corrections have been compared to their CCSD(T) counterparts, while, for larger systems, the quality of the calculated frequencies has been evaluated by comparison with experimental data. It has been shown that B2PLYP yields harmonic frequencies substantially more accurate than other approaches rooted in the density functional theory, and in this respect, it outperforms the B3LYP functional. However, such an improved accuracy is achieved at a significantly increased computational cost, caused by the second-order perturbation treatment of the electron correlation and the slower convergence with respect to the basis set. Nevertheless, when high quality harmonic contributions are required, the availability of the B2PLYP analytic second derivatives shall improve the current state-of-the-art accuracy for significantly larger systems. In addition to accurate harmonic frequencies, the numerical differentiation of the B2PLYP analytic second derivatives provides also cubic and semidiagonal quartic force fields of good quality. However, in this case, despite the significantly larger computational cost, no clear improvement over calculations employing anharmonic force constants obtained at the B3LYP level has been observed. Additionally, in this work, it has been further confirmed that some of the otherwise successful last generation functionals (the M06 and ω B97X families) do not provide sufficiently accurate vibrational properties, concerning both harmonic frequencies and anharmonic contributions. For such reasons, it seems that the most cost-effective approach is currently to add anharmonic corrections calculated at the B3LYP level to harmonic force fields obtained using more sophisticated computational models, like, e.g., CCSD(T) or B2PLYP with large basis sets. In this respect, the B2PLYP/AVTZ//B3LYP/N07D approach combines the feasibility of accurate harmonic frequency computations with the possibility of taking into account the vibrational effects

beyond the harmonic approximation even for quite large systems of biological and/or technological interest.

Acknowledgment. This work was supported by MIUR (PRIN 2006), CNR (PROMO 2006), and Gaussian, Inc. The large scale computer facilities of the VILLAGE network (<http://village.unina.it>) and the Wroclaw Centre for Networking and Supercomputing are acknowledged for providing computer resources.

References

- (1) Jensen, P.; Bunker, P. R. *Computational Molecular Spectroscopy*; John Wiley & Sons: United Kingdom, 2000.
- (2) Carter, S.; Handy, N. C.; Puzzarini, C.; Tarroni, R.; Palmieri, P. *Mol. Phys.* **2000**, *98*, 1697–1712.
- (3) Biczysko, M.; Tarroni, R.; Carter, S. *J. Chem. Phys.* **2003**, *119*, 4197–4203.
- (4) Puzzarini, C.; Barone, V. *Chem. Phys. Lett.* **2008**, *462*, 49–52.
- (5) Puzzarini, C.; Barone, V. *Chem. Phys. Lett.* **2009**, *467*, 276–280.
- (6) Mills, I. M. *Molecular Spectroscopy: Modern Research*; Academic: New York, 1972.
- (7) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108/1–10.
- (8) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.
- (9) East, A. L. L.; Allen, W. D.; Klippenstein, S. J. *J. Chem. Phys.* **1995**, *102*, 8506–8532.
- (10) Martin, J. M. L.; Lee, T. J.; Taylor, P. R.; Francois, J.-P. *J. Chem. Phys.* **1995**, *103*, 2589–2602.
- (11) Dateo, C. E.; Lee, T. J. *Spectrochim. Acta, Part A* **1997**, *53*, 1065–1077.
- (12) Breidung, J.; Thiel, W. *Theor. Chem. Acc.* **1998**, *100*, 183–190.
- (13) Stanton, J. F.; Lopreore, C. L.; Gauss, J. *J. Chem. Phys.* **1998**, *108*, 7190–7196.
- (14) Stanton, J. F.; Gauss, J. *J. Chem. Phys.* **1998**, *108*, 9218–9220.
- (15) Breidung, J.; Thiel, W.; Gaus, J.; Stanton, J. F. *J. Chem. Phys.* **1999**, *110*, 3687–3696.
- (16) Ruden, T.; Taylor, P. R.; Helgaker, T. *J. Chem. Phys.* **2003**, *119*, 1951–1960.
- (17) Puzzarini, C. *J. Chem. Phys.* **2005**, *123*, 024313/1–14.
- (18) Bizzocchi, L.; Degli Esposti, C.; Puzzarini, C. *Mol. Phys.* **2006**, *104*, 2627–2640.
- (19) Puzzarini, C. *J. Mol. Spectrosc.* **2007**, *242*, 70–75.
- (20) Baldacci, A.; Stoppa, P.; Pietropolli Charmet, A.; Giorgianni, S.; Cazzoli, G.; Puzzarini, C. W. L. C. *J. Phys. Chem. A* **2007**, *111*, 7090–7097.
- (21) Tew, D. P.; Klopper, W.; Heckert, M.; Gauss, J. *J. Phys. Chem. A* **2007**, *111*, 11242–11248.
- (22) Barone, V. *J. Phys. Chem. A* **2004**, *108*, 4146–4150.
- (23) Barone, V. *Chem. Phys. Lett.* **2004**, *383*, 528–532.
- (24) Burcl, R.; Handy, N. C.; Carter, S. *Spectrochim. Acta, Part A* **2003**, *59*, 1881–1893.
- (25) Boese, A. D.; Martin, J. J. *J. Phys. Chem. A* **2004**, *108*, 3085–3096.
- (26) Cane, E.; Miani, A.; Trombetti, A. *J. Phys. Chem. A* **2007**, *111*, 8218–8222.
- (27) Cane, E.; Trombetti, A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2428–2432.
- (28) Biczysko, M.; Panek, P.; Barone, V. *Chem. Phys. Lett.* **2009**, *475*, 105–110.
- (29) Puzzarini, C.; Biczysko, M.; Barone, V. *J. Chem. Theory Comput.* **2010**, *6*, 828–838.
- (30) Barone, V.; Cimino, P.; Stendardo, E. *J. Chem. Theory Comput.* **2008**, *4*, 751–764.
- (31) Barone, V.; Cimino, P. *Chem. Phys. Lett.* **2008**, *454*, 139–143.
- (32) Barone, V.; Cimino, P. *J. Chem. Theory Comput.* **2009**, *5*, 192–199.
- (33) Becke, D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (34) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (35) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (36) Barone, V.; Bloino, J.; Biczysko, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1092–1101.
- (37) Barone, V.; Biczysko, M.; Cimino, P. Interplay of stereo electronic vibrational and environmental effects in tuning physico-chemical properties of carbon centered radicals. In *Carbon-Centered Free Radicals and Radical Cations*; Forbes, M. D. E., Ed.; John Wiley & Sons, Inc.: New York, 2010; pp 105–139.
- (38) Bloino, J.; Biczysko, M.; Santoro, F.; Barone, V. *J. Chem. Theory Comput.* **2010**, *6*, 1256–1274.
- (39) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108/1–16.
- (40) Grimme, S.; Neese, F. *J. Chem. Phys.* **2007**, *127*, 154116/1–18.
- (41) Grimme, S.; Goerigk, L. *J. Phys. Chem. A* **2009**, *113*, 767–776.
- (42) Schneider, W.; Thiel, W. *Chem. Phys. Lett.* **1989**, *157*, 367–373.
- (43) Stanton, J. F.; Gauss, J. *Int. Rev. Phys. Chem.* **2000**, *19*, 61–95.
- (44) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.
- (45) Johnson, B.; Frisch, M. *J. Chem. Phys.* **1994**, *100*, 7429–7442.
- (46) Johnson, B.; Frisch, M. *Chem. Phys. Lett.* **1993**, *216*, 133–140.
- (47) Stratmann, R.; Burant, J.; Scuseria, G.; Frisch, M. *J. Chem. Phys.* **1997**, *106*, 10175–10183.
- (48) Pople, J.; Krishnan, R.; Schlegel, H.; Binkley, J. *Int. J. Quantum Chem.* **1979**, *16-S13*, 225–241.
- (49) Frish, M. J.; Head-Gordon, M.; Pople, J. *Chem. Phys. Lett.* **1990**, *166*, 275–280.
- (50) Frish, M. J.; Head-Gordon, M.; Pople, J. *Chem. Phys. Lett.* **1990**, *166*, 281–289.
- (51) Gauss, J.; Stanton, J.; Bartlett, R. *Chem. Phys. Lett.* **1992**, *195*, 194–199.
- (52) Gauss, J.; Stanton, J.; Bartlett, R. *J. Chem. Phys.* **1992**, *97*, 7825–7828.

- (53) Cammi, R.; Mennucci, B.; Pomelli, C.; Cappelli, C.; Corni, S.; Frediani, L.; Trucks, G.; Frisch, M. *Theor. Chim. Acta* **2004**, *111*, 66–77.
- (54) Neese, F.; Schwabe, T.; Grimme, S. *J. Chem. Phys.* **2007**, *126*, 124115/1–15.
- (55) Scalmani, G.; Frisch, M. J.; Mennucci, B.; Tomasi, J.; Cammi, R.; Barone, V. *J. Chem. Phys.* **2006**, *124*, 094107/1–15.
- (56) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3093.
- (57) Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2010**, *132*, 114110/1–15.
- (58) Frish, M. J.; Head-Gordon, M.; Pople, J. *Chem. Phys.* **1990**, *141*, 189–196.
- (59) Double- and triple- ζ basis sets of the N07 family are available for download; visit <http://idea.sns.it> (accessed April 17, 2010).
- (60) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (61) Kendall, A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.
- (62) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Burant, J.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fo, D. J. *Gaussian Development Version*, revision H.08; Gaussian, Inc.: Wallingford, CT, 2009.
- (63) Barone, V.; Festa, G.; Grandi, A.; Rega, N.; Sanna, N. *Chem. Phys. Lett.* **2004**, *388*, 279–283.
- (64) Begue, D.; Carbonniere, P.; Pouchan, C. *J. Phys. Chem. A* **2005**, *109*, 4611–4616.
- (65) Carbonniere, P.; Lucca, T.; Pouchan, C.; Rega, N.; Barone, V. *J. Comput. Chem.* **2005**, *26*, 384–388.
- (66) Puzzarini, C.; Barone, V. *J. Chem. Phys.* **2008**, *129*, 084306/1–7.
- (67) Puzzarini, C.; Barone, V. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6991–6997.
- (68) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fo, D. J. *Gaussian 09*, revision A.02; Gaussian Inc.: Wallingford, CT, 2009.
- (69) Zhao, Y.; Schults, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.
- (70) Zhao, Y.; Truhlar, D. G. *Theor. Chim. Acta* **2008**, *120*, 215–241.
- (71) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106/1–15.
- (72) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.
- (73) Henderson, T.; Izmaylov, A. F.; Scalmani, G.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 044108/1–9.
- (74) Jacquemin, D.; Perpète, E.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105/1–12.
- (75) Hamprecht, F. A.; Cohen, A.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264–6271.
- (76) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101/1–14.
- (77) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849–1868.
- (78) Martin, J. M. L.; El-Yazal, J.; Francois, J.-P. *Mol. Phys.* **1995**, *86*, 1437.
- (79) NIST Chemistry Webbook. <http://webbook.nist.gov/chemistry> (accessed Dec 9, 2006).
- (80) Martin, J. M. L. *J. Chem. Phys.* **1994**, *100*, 8186–8193.
- (81) Feller, D.; Peterson, K. *J. Chem. Phys.* **2009**, *131*, 154306/1–10.
- (82) Puzzarini, C. *Theor. Chim. Acta* **2008**, *121*, 1–10.
- (83) Rauhut, G.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054105/1–10.
- (84) Barone, V. *J. Phys. Chem. A* **2004**, *108*, 4146–4150.
- (85) O'Maley, P. *THEOCHEM* **2005**, *755*, 147–150.
- (86) Bloino, J.; Biczysko, M.; Crescenzi, O.; Barone, V. *J. Chem. Phys.* **2008**, *128*, 244105/1–15.
- (87) Marenich, A. V.; Boggs, J. E. *J. Phys. Chem. A* **2003**, *107*, 2343–2350.
- (88) Puzzarini, C. *Theor. Chem. Acc.* **2008**, *120*, 325–336.
- (89) Sharma, A. R.; Braams, B. J.; Carter, S.; Shepler, B. C.; Bowman, J. M. *J. Chem. Phys.* **2009**, *130*, 174301/1–9.
- (90) Klots, T. D. *Spectrochim. Acta, Part A* **1998**, *54*, 1481–1498.
- (91) Mellouki, A.; Lievin, J.; Herman, M. *Chem. Phys.* **2001**, *271*, 239–266.
- (92) Klots, T. D.; Chirico, R. D.; Steele, W. V. *Spectrochim. Acta, Part A* **1994**, *50*, 765–795.
- (93) Graindourze, M.; Smets, J.; Zeegers-Huyskens, T.; Maes, G. *J. Mol. Struct.* **1990**, *222*, 345–364.
- (94) Roth, W.; Imhof, P.; Gerhards, M.; Schumm, S.; Kleiner-manns, K. *Chem. Phys.* **2000**, *252*, 247–256.

Formation Enthalpies of Ions: Routine Prediction Using Atom Equivalents

Didier Mathieu* and Yohann Pipeau

CEA, DAM, Le Ripault, F-37260 Monts, France

Received January 13, 2010

Abstract: In view of identifying routine procedures to estimate formation enthalpies of ionic systems such as energetic salts or ionic liquids on the basis of density functional theory (DFT), various combinations of atom equivalent (AE) schemes, functionals, and basis sets are compared, using a specially designed training set to parametrize the models. After correction, none of the functionals considered proves significantly more reliable than B3LYP. A small but systematic improvement is noted as AE values are allowed to depend on the atomic environment. However, AE parameters fail to make up for basis set limitations, in contrast to previous observations for neutrals. Finally, a good trade-off between reliability and cost is obtained for ions using B3LYP/6-31++G** energies.

1. Introduction

In recent years, molecular ionic species have received much interest as constituents of ionic liquids or high energy density materials.^{1–3} Depending on the application in view, ions with high stability and/or high energy content are needed. In this context, procedures to predict the formation enthalpies of ionic compounds are needed to focus new syntheses on the most promising candidates. While a plethora of methods are available for neutral molecules, such as group additivity schemes,⁴ molecular mechanics,^{5,6} or quantitative structure property relationships,^{7,8} evaluation of the standard formation enthalpy $\Delta_f H^\circ$ of ions in gas-phase must resort to quantum mechanics for most cases of practical interest.^{9–12} Provided a suitable basis set is used, high level *ab initio* methods perform well for charged species without any specific treatment.^{13,14}

However, more efficient methods are needed to predict $\Delta_f H^\circ$ for arbitrary ions on a routine basis. Unfortunately, detailed investigations of the performance of such procedures when applied to charged species are still lacking. Available data indicate that semiempirical methods are not suitable for such systems,^{15,16} while simple approaches based on density functional theory (DFT) yield reasonable estimates of the energy content of energetic salts.^{17,18} Therefore, this work compares simple quantum procedures, with a focus on

efficient DFT-based methods involving empirical local corrections. Such procedures have been extensively investigated for neutrals.^{19,20} However, they might be less effective for ions in view of the delocalization arising from the fact that their wave function usually involves several canonical structures.

More specifically, the present work examines how the performance of standard correction schemes is affected on going from neutrals to ions. Experimental gas phase enthalpies of ions, taken from the literature, are compared to the results of extensive calculations using various levels of theory for the electronic energy, different procedures to convert this energy into formation enthalpies, and a newly introduced approach to define the training set employed to fit the adjustable parameters involved. A clear picture of the accuracy that may be obtained according to the specific procedure employed is thus obtained. Cost effective procedures are identified, with a reliability consistent with the uncertainties associated with the evaluation of intermolecular interactions.^{9,18,21} They are now implemented in a user-friendly software package²² and allow bench chemists to perform routine calculations of $\Delta_f H^\circ$ for materials of potential interest. On the other hand, in contrast to many recent papers in the field,^{23–25} no attempt is made to improve the accuracy of state-of-the-art DFT methodologies for gas-phase enthalpies.

* To whom correspondence should be addressed. E-mail: didier.mathieu@cea.fr.

2. Computational Methods

2.1. Theoretical Framework. By definition, the formation enthalpy of a given compound made of different atomic elements H, C, N, ... with $Z = 1, 6, 7, \dots$ is obtained as the difference between the theoretical enthalpy H° of the compound and the corresponding enthalpies $H^\circ(Z, \text{stp})$ of every atom Z in its standard reference state (i.e., standard temperature and pressure):

$$\Delta_f H^\circ = H^\circ - \sum_Z n_Z H^\circ(Z, \text{stp}) \quad (1)$$

Here, n_Z is the number of atoms with atomic number Z in the compound. For most light elements, the theoretical evaluation of $H^\circ(Z, \text{stp})$ is straightforward. For instance, for the nitrogen atom, it is simply $H^\circ(\text{N}_2, \text{stp})/2$ since its standard reference state is the N_2 molecule. Nevertheless, in some cases, $H^\circ(Z, \text{stp})$ may be more difficult to calculate. Even for carbon, whose reference state (graphite) is well-known, a problem arises because many theoretical methods are only implemented for finite systems and cannot be used to compute $H^\circ(\text{C}_{\text{graphite}}, \text{stp})$.

This problem is avoided by the use of molecular or atomic reference states, i.e., gaseous species i with well-established formation enthalpies $\Delta_f H^\circ(i)$. The formation enthalpy is then obtained as

$$\Delta_f H^\circ = \sum_i N_i \Delta_f H^\circ(i) + (H^\circ - \sum_i N_i H^\circ(i)) \quad (2)$$

where the numbers N_i of the reference species must be consistent with the empirical formula of the compound studied. In other words, if n_Z^i stands for the number of atoms Z in species i , the values of N_i must satisfy for every element Z :

$$\sum_i n_Z^i N_i = n_Z \quad (3)$$

In principle, any set of reference compounds i can be used as long as the linear system defined by eq 3 is regular and exhibits a solution $\{N_i\}$. Notwithstanding the role of the level of theory used to compute H° data, it is important to keep in mind that calculated $\Delta_f H^\circ$ values also depend in practice on the reference compounds selected.

In some cases, reference compounds i with accurate $\Delta_f H^\circ$ data available may be found in such a way that the number of each kind of chemical bond is the same for the reference compounds and for the molecule studied. In other words, the compound studied is obtained from species i through an isodesmic reaction.⁴ This situation is especially favorable as the calculation of the enthalpy difference between the molecule under study and reference species benefits from an effective cancellation of errors. In principle, a computer algorithm coupled with a database for accurate $\Delta_f H^\circ$ data could be used to identify suitable reference species for every new compound studied. In the lack of such a program, the use of isodesmic reactions is not convenient for routine calculations. In fact, it is not always possible to identify a suitable isodesmic scheme. For ions, this approach is

problematic as bond orders cannot always be assigned unambiguously.

Relaxing the constraint to rely on isodesmic reactions allows more practical approaches. For instance, a molecular reference species Z_m may be introduced for every atomic element Z in the compound studied.²⁶ Possible values of m are $m = 2, 60, 2, 2, 8, \dots$ for elements H, C, N, O, S, ..., respectively. In this case, the linear system defined by eq 3 is diagonal. As a result, eq 2 becomes simply

$$\Delta_f H^\circ = H^\circ + \sum_Z \frac{n_Z}{m} (\Delta_f H^\circ(Z_m) - H^\circ(Z_m)) \quad (4)$$

It is clear from eq 4 that the role of the differences $\Delta_f H^\circ(Z_m) - H^\circ(Z_m)$ between experimental and theoretical enthalpies is simply to shift the zero of enthalpies in order to make theoretical data consistent with the conventional thermodynamic reference state. On the other hand, the second term in this equation depends in practice on the reference compounds used because of the uncertainties associated with experimental $\Delta_f H^\circ(Z_m)$ and theoretical $H^\circ(Z_m)$ data.

As the formation enthalpies of gaseous atoms are well-known for the main group elements, their use as reference states yields small uncertainties associated with experimental data. Formation enthalpies are then obtained as the difference between the formation enthalpies of the gaseous atoms and the atomization enthalpy of the compound:

$$\Delta_f H^\circ = \sum_Z n_Z \Delta_f H^\circ(Z) - (\sum_Z n_Z H^\circ(Z) - H^\circ) \quad (5)$$

Unfortunately, this approach leads to much larger errors than isodesmic reaction schemes since the errors on calculated enthalpies do not effectively cancel for atomization reactions. A simple approach to remove these systematic errors consists in introducing empirical parameters X_Z :

$$\Delta_f H^\circ = H^\circ + \sum_Z n_Z C_Z = H^\circ + \sum_Z n_Z (\Delta_f H^\circ(Z) - H^\circ(Z)) + \sum_Z n_Z X_Z \quad (6)$$

In further attempts to improve the results, X_Z may be assumed to depend not only on the atomic number Z of the atom but also on its environment within the molecule under study:

$$\Delta_f H^\circ = H^\circ + \sum_Z n_Z (\Delta_f H^\circ(Z) - H^\circ(Z)) + \sum_A X_A \quad (7)$$

The last sum in eq 7 runs over every atom in the compound studied. The present work compares the performance of procedures based on this equation, according to the level of theory used to compute H° and the actual definition adopted for the X_A parameters, hereafter referred to as atom equivalents (AEs).

Beyond AE schemes, a number of alternatives have been introduced to convert HF or DFT energies into formation enthalpies while correcting their main deficiencies.^{27,28} These methods focus on neutral molecules and rely on empirical relationships. However, simpler procedures based on eq 7 are not necessarily less reliable.^{29,39} In favorable situations,

Table 1. Definition of the Correction Procedures P3, P5, P8, and P10 Used in This Work

P3	specific AE for H, $X_A = a + bZ$ for $A = \text{C, N, O, F}$
P5	one AE for every element: H, C, N, O, F
P8	additional AEs C' , N' , O' for atoms with multiple bonds
P10	H1, C4, C3, C2, N3, N2, N1, O2, O1, F1

they yield average absolute deviations (AAD) from experimental results close to 2.5 kcal/mol.^{30,40} Therefore, all models considered in the present work rely on eq 7. They differ only in the definition adopted for the X_A parameters.

2.2. Definition of Atom Equivalents. Five procedures are considered in this work for predicting formation enthalpies of ions from theoretical total energies and eq 7. They are named P_m where m is the number of adjustable parameters required to fit $\Delta_f H^\circ$ values of compounds made of HCNOF atoms. In procedure P5, X_A is assumed to be independent of the environment of the atom. It depends only on the atomic number Z of the atom. Therefore, only five adjustable parameters are required for HCNOF. Procedure P8 introduces three additional AEs C' , N' , and O' for atoms involved in multiple bonds, as done by Rice and co-workers.^{39,41}

In procedure P10, an atom equivalent depends not only on the atomic symbol X but also on the atom coordination number n .⁴² It is thus denoted Xn . In principle, the 10 P10 parameters listed in Table 1 do not allow for handling of ammonium and hydronium cations, due to the lack of N4 and O3 parameters. In fact, it was previously noted that using the N3 value instead of N4 for ammonium salts yields satisfactory results.¹⁷ In this work, using N3 and O2 parameters for NH_4^+ and H_3O^+ does not lead to specially large deviations. This indicates that N4 and O3 values derived from these two ions are close to the N3 and O2 parameters, which refer to atoms in different bonding environments but with the same hybridizations. The derivation of optimal N4 and O3 values is beyond the scope of this work, as such values should be averaged over typical atomic environments. Moreover, for reasons detailed in section 3, present parameters are derived from gas-phase data for neutrals.

Finally, procedure P3 aims at reducing the number of adjustable parameters by taking advantage of the linear correlation often observed between the values of the equivalents for CNOF atoms and the corresponding atomic numbers. In other words, while a constant value is attributed to the H equivalent, the others are assumed to vary according to $X_A = a + bZ$ where Z is the atomic number of atom A . This leaves only three adjustable parameters: a , b , and H . Such a procedure to reduce the number of empirical parameters is especially interesting in view of extending AE methods beyond first-row atoms.

These procedures are summarized in Table 1. In earlier studies, specific AEs or group equivalents are sometimes introduced for some special chemical groups, such as nitros or azides.^{41,42} However, such group specific parameters hamper the generality of the procedure. In fact, they are often unsuitable for ions because of ill-defined bond orders. Accordingly, no attempt is made here to introduce such group parameters. On the other hand, alternative approaches based

on bond equivalents (BEs), charge-dependent AEs, or BEs depending on Mulliken bond populations have also been investigated. However, for the present compounds, they prove significantly worse than the P_n procedures described above, despite encouraging results sometimes reported for other data sets or theoretical levels.^{30,31,43}

2.3. Computational Procedures. Having defined the atom equivalents X_A involved in eq 7, computational procedures remain to be selected to derive theoretical enthalpies $H^\circ(Z)$ for gaseous atoms and H° for the compound under study. In this work, $H^\circ(Z)$ is computed at the G3MP2B3 level,⁴⁴ a specially efficient version of the well-known G3 composite method.⁴⁵ This choice is irrelevant as any deficiency in $H^\circ(Z)$ will be absorbed into the AE values X_A . However, an explicit evaluation of $H^\circ(Z)$ with reasonable accuracy makes it possible to interpret X_A as approximate corrections to H° .

The evaluation of molecular enthalpies H° from total electronic energies E_0 (frozen atoms at 0 K) and vibrational frequencies is straightforward within the ideal gas and harmonic approximations. Because frequencies add a significant computational overhead compared with single-point energy calculations, simple additive schemes have been developed to estimate $H^\circ - E_0$, which includes the zero-point energy as well as thermal contributions.^{30,43,46} In this work, it is obtained from standard enthalpic corrections $H_{\text{corr}}(Z)$ introduced by Winget and Clark:⁴⁶

$$H^\circ = E_0 + \sum_Z n_Z H_{\text{corr}}(Z) \quad (8)$$

Finally, the approach based on eqs 7 and 8 amounts to adding atomic parameters to the total quantum chemical energy E_0 in order to obtain the formation enthalpy:

$$\Delta_f H^\circ = E_0 + \sum_Z n_Z (\Delta_f H^\circ(Z) - H^\circ(Z) + H_{\text{corr}}(Z)) + \sum_A X_A = E_0 + \sum_A Y_A \quad (9)$$

The Y_A parameters introduced previously⁴² and also referred to as $-\epsilon_A$ ¹⁸ are commonly used for straightforward conversion of E_0 data into $\Delta_f H^\circ$ values. In this work, it was decided to make their various contributions explicit. This facilitates the interpretation of the X_A parameters. Since the latter should ideally be zero, their magnitude provides an estimate of the errors in the other contributions. Previous studies show that these errors are dominated by the uncertainties associated with E_0 values.^{42,46}

The derivation of formation enthalpies from eq 7 involves the calculation of total energies E_0 . They are computed using various levels of theory: the nonlocal exchange HF functional;⁴⁷ the self-consistent-charge density functional tight binding scheme (SCC-DFTB);⁴⁸ functionals based on the local density approximation (LDA): $X\alpha$ ⁴⁹ and SVWN;^{49,50} functionals based on the generalized gradient approximation (GGA): BP86,^{51,52} BLYP,^{51,53,54} PW91,^{55,56} mPW91,⁵⁷ PBE,^{58,59} and HCTH;⁶⁰ hybrid GGA functionals (H-GGA): B3LYP,^{51,53,61} B3P86,⁵¹⁻⁵³ B3PW91,^{51,53,55} PBE1PBE,⁶² B1LYP,⁶³ B98,⁶⁴ and the ‘‘half and half’’ functionals

BHandH and BHandHLYP implemented in Gaussian⁶⁵ following those introduced by Becke,⁶⁶ meta GGA functionals (M-GGA): TPSS⁶⁷ and VSXC;⁶⁸ and hybrid meta GGA functionals (HM-GGA): B1B95⁶¹ and BMK.⁶⁹

The large number of functionals presently considered stems from the fact that AE-based correction procedures have been so far applied only to a few popular functionals, especially BP86 and B3LYP. A systematic application of such procedures is of interest because functionals discarded as inaccurate on the basis of raw $\Delta_f H^\circ$ predictions might prove valuable if the errors lend themselves to effective corrections. For instance, earlier attempts to obtain good thermochemistry from $X\alpha$ calculations rely on atom-dependent values of the Slater exchange parameter α rather than AE-based corrections.⁷⁰ Although so far unsuccessful, the search for effective procedures to estimate $\Delta_f H^\circ$ from $X\alpha$ calculations is of special interest as the $X\alpha$ functional lends itself to a fully analytic calculation of the Hamiltonian matrix, in contrast to the others.⁷¹ In this work, the standard value of 0.7 is used for the Slater coefficient in order to make up for the lack of an explicit correlation model. With regard to the correlation part of SVWN, the default version in Gaussian⁶⁵ is used, namely, the one numbered III in the original paper.⁵⁰ Present nomenclature for other DFT functionals follows the one adopted in a recent review of their performance, in which a more comprehensive list of references may be found.⁷² More recent HM-GGA functionals, including TPSSh⁶⁷ or the M06 family,^{25,73} are not yet available in our group and lie beyond the scope of the present paper. Although they are relatively costly, these functionals perform remarkably well without *a posteriori* corrections. It will be of interest to investigate whether their predictions can be further enhanced by correction procedures such as those considered here.

In addition to the new AEs introduced in the present paper, alternative approaches to $\Delta_f H^\circ$ are used for comparison: G3MP2B3,⁴⁴ and popular semiempirical methods based on the NDDO approximation, namely, AM1,⁷⁴ PM3,^{75,76} and RM1.¹⁵ The latter is a recent reparametrization of the AM1 Hamiltonian which provides a remarkable improvement for the prediction of formation enthalpies of organic and biological molecules. In order to compare the predictive power of the models for ions, the root-mean-square deviation (RMSD) between calculated and observed values is used as the main criterion. However, the average absolute deviation (AAD) is also reported to make comparison with previous work easier. On the other hand, although this study focuses on ions, RMSD and AAD values derived from a leave-one-out cross validation of the training set data are also considered as rough indicators of the reliability of these procedures for neutral systems. The following software is used for all present calculations: MOPAC7 for semiempirical methods,⁷⁷ the original DFTB code for SCC-DFTB,⁴⁸ and Gaussian 03W for *ab initio* and DFT methods.⁶⁵

3. Database

The selection of a suitable database to assess or parametrize computational procedures is no trivial task. The scope of the method, its expected accuracy, or the number of

adjustable parameters to be fitted must be considered. Over the years, highly accurate thermochemical data have been collected to assess the performance of high-level theories, especially composite *ab initio* models.^{78–82} Since they exhibit only relatively small species, such data sets are not optimal to parametrize more approximate procedures applicable to large organic compounds.

Such procedures, based on either molecular mechanics, semiempirical Hamiltonians, HF, or DFT are developed using larger data sets obtained by including somewhat less reliable data, often without error bars.^{15,16,83,84} This is acceptable, as the corresponding procedures yield typical errors significantly larger than experimental uncertainties. Extended data sets used to develop general schemes invariably exhibit a significant proportion of hydrocarbons or other simple, monofunctionalized organic compounds. As a result, they might provide too optimistic views of the reliability of a given procedure when applied to unusual compounds, such as molecular ions.⁴⁶

An alternative approach consists in developing specific parameters on the basis of a restricted family of compounds. For instance, training sets focused on nitro compounds have been used to derive AEs specially optimized for energetic materials.^{39,41} Specialized equivalents have also been published for hydrocarbons,^{85,86} propellanes,⁸⁷ and some monofunctionalized compounds.^{88,89} For molecular ions, this strategy is not well suited owing to the scarcity and relative lack of reliability of available gas phase data. Furthermore, while AE values depend on local atom environments, they should not depend on the total charge of the compound. Indeed, besides a charged group, ions may exhibit the same functional groups as neutral molecules. Therefore, it would make no sense to develop specific AEs for ionic systems.

Accordingly, present procedures are parametrized exclusively against data for neutral CHNOF molecules compiled in Table 2, while $\Delta_f H^\circ$ data for ions (Tables 3 and 4) are used only for validation purposes. This provides a stringent test of the transferability of the parameters. For most compounds in the training set, $\Delta_f H^\circ$ is reported to within <1 kcal/mol (Table 2). In contrast, error bars are often unavailable for ions. Therefore, $\Delta_f H^\circ$ data reported in Tables 3 and 4 are prone to large uncertainties, and one should not attach too much significance to individual values. Nevertheless, they are significantly more reliable on average than present DFT-based values, as confirmed by their overall good agreement with G3MP3B3 data (Tables 3 and 4). Therefore, the corresponding RMSD and AAD values provide suitable comparison criteria to assess the performance of the present procedures.

The present training set is specially designed to ensure a balanced coverage of the many possible chemical environments for an atom in polyatomic species, thus avoiding a bias of the parametrization due to the prevalence of specific moieties such as alkyl groups. First, all possible bonds between CNOF atoms are listed, considering only 1, 2, and 3 as possible formal values for the bond orders. Then, dangling bonds are saturated with H atoms. This yields 24 simple compounds for which experimental $\Delta_f H^\circ$ values are available. For each of these compounds, and whenever

Table 2. Experimental Formation Enthalpies (with Error Bars when Available) and Corresponding Deviations for Theoretical Values Calculated Using G3MP2B3, PM3, and DFT, More Specifically the P5 Procedure Applied to B3LYP/6-31++G** Total Energies^a

compound	CAS number	exptl.	G3MP2B3	PM3	DFT
HN=NH (<i>trans</i>)	15626-43-4	50.9 ± 2	-2.9	-13.1	-3.3
C ₁₀ H ₈ (naphthalene)	91-20-3	35.9 ± 2	-2.6	4.8	1.0
H ₅ C ₆ -NO ₂	98-95-3	16.4 ± 0.2	-2.2	-1.7	-1.2
C ₄ H ₄ N ₂	289-95-2	46.8 ± 0.4	-2.2	-8.6	-4.8
F-CH=CH ₂	75-02-5	-32.5	-1.9	3.8	-3.1
C ₉ H ₇ N (quinoline)	91-22-5	47.9	-1.9	-0.2	-0.2
H ₅ C ₆ -N=N-C ₆ H ₅	17082-12-1	96.9 ± 0.3	-1.9	-6.0	-0.7
HCN	74-90-8	32.3	-1.7	0.5	-1.9
(CH ₃) ₂ C=O	67-64-1	-49.8 ± 0.1	-1.4	-3.3	-2.6
H ₂	1333-74-0	0.0	-1.0	-13.4	6.0
C ₆ H ₆	71-43-2	19.8 ± 0.2	-0.7	3.6	-0.5
C ₅ H ₅ N	110-86-1	33.6	-0.7	-3.1	-1.9
H ₂ C=O	50-00-0	-26.0 ± 0.1	-0.7	-8.1	0.7
C ₂ H ₄ N ₄ (1-methyl-1H-tetrazole)	16681-77-9	77.2 ± 0.5	-0.5	6.9	-1.0
H ₂ C=CH ₂	74-85-1	12.5 ± 0.1	-0.5	4.1	0.7
HCCCH ₃	74-99-7	44.3 ± 0.2	-0.5	-4.1	-1.0
HCCH	74-86-2	54.3 ± 0.2	-0.2	-3.6	1.4
H ₅ C ₆ -NH ₂	62-53-3	20.8 ± 0.2	-0.2	0.5	0.5
H ₃ CF	593-53-3	-56.0	-0.2	2.2	-3.1
HF	7664-39-3	-65.3 ± 0.2	0.0	2.4	1.4
C ₃ H ₃ NO (isoxazole)	288-14-2	19.6 ± 0.1	0.0	15.3	-0.2
H ₂ C=CH-CH ₂ -CH ₃	106-98-9	-0.1 ± 0.2	0.0	1.4	0.2
H ₂ C=C(CH ₃) ₂	115-11-7	-4.3 ± 0.3	0.2	0.7	1.0
C ₄ H ₄ N ₂ (pyridazine)	289-80-5	66.5 ± 0.3	0.2	-10.5	-1.4
H ₃ C-CH ₂ -OH	64-17-5	-55.9 ± 0.5	0.2	-0.7	-1.7
H ₂ O	7732-18-5	-57.8 ± 0.01	0.2	4.3	2.4
H ₃ C-CH ₃	74-84-0	-20.0 ± 0.1	0.2	1.7	-1.4
CH ₄	74-82-8	-17.9	0.2	4.8	0.2
H ₃ C-OH	67-56-1	-48.1 ± 3	0.2	-3.6	-1.0
NF ₃	7783-54-2	-31.6	0.2	7.2	-2.6
H ₃ C-CH ₂ -CH ₂ -CH ₃	106-97-8	-30.0 ± 0.2	0.5	1.0	-1.4
H ₃ C-NO ₂	75-52-5	-17.9	0.5	1.9	-0.5
(CH ₃) ₃ CONO	540-80-7	-41.1 ± 1	0.5	16.0	2.6
HO-N=O	7782-77-6	-18.3	0.5	4.8	0.7
N ₂	7727-37-9	0.0	0.5	17.4	0.5
F ₃ C-NF ₂	335-01-3	-169.0 ± 0.6	0.5	1.0	-1.0
NH ₃	7664-41-7	-11.0	0.7	-1.9	0.2
C ₈ H ₆ N ₂ (phthalazine)	253-52-1	78.8 ± 0.8	0.7	-6.5	1.9
H ₂ N-C ₆ H ₄ -NO ₂ (<i>p</i> -nitroaniline)	100-01-6	13.2 ± 0.4	0.7	-2.4	0.7
C ₄ H ₈ N ₂ O ₃ (4-nitromorpholine)	4164-32-3	-31.3 ± 0.4	0.7	-1.7	-1.0
H ₃ C-ONO ₂	598-58-3	-29.1 ± 0.3	0.7	-3.1	-1.9
H ₃ C-CH ₂ -CH ₂ -NH ₂	107-10-8	-16.7 ± 0.2	0.7	0.0	-0.2
(CH ₃) ₃ C-NO ₂	594-70-7	-42.3 ± 0.8	1.0	10.0	3.6
H ₃ C-COOH	64-19-7	-103.5 ± 0.6	1.0	1.4	-0.7
C ₂ H ₃ N ₃ (1,2,4-triazole)	288-88-0	46.1 ± 0.2	1.0	5.7	-0.2
HOOH	7722-84-1	-32.5	1.2	-8.1	0.5
H ₃ C-NH ₂	74-89-5	-5.5	1.2	0.2	0.2
HO-NO ₂	7697-37-2	-32.1	1.2	-5.7	-0.5
F ₂	7782-41-4	0.0	1.4	-21.5	-0.0
(CH ₃) ₂ N-NO ₂	4164-28-7	-1.2 ± 0.3	1.4	2.4	-1.4
OF ₂	7783-41-7	5.9	1.4	-10.5	-2.4
H ₃ C-OOH	3031-73-0	-31.3	1.7	-5.7	-1.0
HN=O	14332-28-6	23.8	1.7	-9.8	4.8
H ₂ N-OH	7803-49-8	-11.4	1.7	-1.9	0.7
FO-NO ₂	7789-26-6	2.5	1.7	-8.4	-2.6
H ₂ N-NH ₂	302-01-2	22.8	1.7	-1.9	1.0
C ₄ H ₄ N ₂ (pyrazine)	290-37-9	46.9 ± 0.4	1.9	-7.4	-0.7
H ₂ C=CH-CN	107-13-1	41.3	2.4	8.8	1.0
C ₆ H ₁₀ N ₂ O ₂ (1-nitropiperidine)	7119-94-0	-10.6 ± 0.6	2.9	3.1	2.2
C ₈ H ₆ N ₂ (quinoxaline)	91-19-0	57.4 ± 0.8	3.3	-1.2	3.3
HOF	14034-79-8	-23.5	3.6	-5.5	2.2
H ₂ C=NH	2053-29-4	16.0 ± 2	4.8	5.0	5.0
FNH ₂	15861-05-9	-11.5 ± 0.6	5.7	6.7	3.8
F ₃ C-OF	373-91-1	-182.8 ± 2.4	6.5	-4.3	4.8

^a Experimental values taken from the NIST Webbook,⁹⁰ except for HN=NH whose formation enthalpy is taken from ref 96. The reader is referred to references therein for further details. Unit: kcal/mol.

possible, a new molecule is obtained by substitution of a hydrogen atom, in such a way that $\Delta_f H^\circ$ is available in the

NIST Webbook for the new derivative.⁹⁰ At this stage, aromatic systems are not represented in the database since

Table 3. Same Data As in Table 2 for Cations^a

compound	exptl.	G3MP2B3	PM3	DFT
C ₃ H ₅ ⁺ (cyclopropyl)	235.0 a	-6.2	-2.2	-9.8
NO ₂ ⁺ (nitrogen dioxide)	233.0 a	-4.8	-24.4	1.9
C ₄ H ₉ ⁺ (isobutyl)	176.0 a	-3.3	2.6	-7.6
OH ⁺ (triplet)	309.1 b	-2.9	-19.4	6.0
NH ₄ ⁺ (ammonium)	155.0 a	-2.2	-1.4	-2.4
CH ⁺	387.8 b	-2.2	-21.7	9.3
CHO ⁺ (formyl)	199.0 a	-1.4	-22.0	4.5
C ₄ H ₇ ⁺ (methyl allyl)	207.9 b	-1.4	3.8	-7.4
NO ⁺ (nitric oxide)	237.0 a	-1.2	1.2	8.4
NH ₂ ⁺ (triplet)	302.0 b	-1.2	-41.8	4.3
C ₂ H ₄ ⁺ (ethylene)	257.0 a	-0.7	-8.1	-5.3
C ₃ H ₅ ⁺ (propenyl)	237.0 a	-0.7	1.2	-4.5
C ₄ H ₅ O ⁺ (C-protonated furan)	165.0 b	-0.5	10.3	-3.3
C ₅ H ₆ N ⁺ (pyridinium)	178.0 b	0.0	9.1	-3.1
CH ₃ O ⁺ (H ₂ COH)	169.3 b	0.7	-2.9	1.2
CH ₃ ⁺ (methyl)	261.0 a	1.0	-4.3	3.3
C ₂ H ₅ ⁺ (ethyl)	216.0 a	1.2	6.5	0.0
CH ₄ N ⁺	179.4 b	1.2	5.7	0.7
C ₃ H ₃ ⁺ (cyclopropenyl)	257.0 a	1.2	12.7	0.2
C ₇ H ₇ ⁺ (tropolium)	209.0 a	1.2	12.0	-4.8
HCNH ⁺	225.8 b	1.7	-12.2	4.5
OCOH ⁺	141.0 b	1.7	-1.4	3.6
CH ₃ CO ⁺	156.0 b	1.9	2.9	2.9
C ₆ H ₅ ⁺ (phenyl)	269.3 b	1.9	19.4	0.7
C ₃ H ₇ ⁺ (1-propyl)	211.0 b	2.2	3.3	-21.5
CH ₄ N ⁺ (methaniminium)	178.0 a	2.6	7.2	2.2
CH ₃ CNH ⁺	195.0 b	2.6	2.6	1.4
C ₃ H ₅ ⁺ (allyl)	226.0 a	2.6	6.7	-1.0
HO-CH-OH ⁺	96.0 b	3.1	-0.7	1.7
CH ₃ OH ₂ ⁺	136.0 b	3.1	20.6	1.2
C ₄ H ₉ ⁺ (<i>n</i> -butyl)	183.0 b	3.1	7.6	-1.9
C ₃ H ₇ ⁺ (2-propyl)	190.9 b	3.3	6.2	-1.4
C ₃ H ₃ ⁺ (propynyl)	281.0 a	3.6	-5.5	0.2
CH ₃ -OH-CH ₃ ⁺	130.0 b	4.3	27.0	0.5
CH ₃ CHOH ⁺	139.0 b	4.5	5.5	1.9
C ₂ H ₃ ⁺ (vinyl)	266.0 a	5.0	-1.9	5.7
C ₄ H ₉ ⁺ (<i>tert</i> -Butyl)	165.8 b	5.3	12.0	1.0
C ₇ H ₇ ⁺ (benzyl)	212.0 a	5.5	15.3	1.9
C ₅ H ₉ ⁺ (cyclopentyl)	188.0 a	5.7	5.3	2.4
H ₃ O ⁺ (hydronium)	138.9 a	6.0	20.1	7.6
C ₄ H ₇ ⁺ (2-butenyl)	200.0 a	6.2	11.7	0.5
C ₆ H ₁₁ ⁺ (cyclohexyl)	177.0 a	7.2	9.1	3.1
C ₄ H ₇ ⁺ (cyclobutyl)	213.0 a	17.9	12.4	13.4

^a Sources: *a* = ref 15, *b* = ref 98.

only integer bond orders have been considered. Therefore, additional aromatic and nitro compounds are subsequently included in order to introduce fractional formal bond orders.

Finally, this work relies on a training set of 64 neutral molecules and a validation set of 73 ions, including 43 cations and 30 anions. All of these compounds are listed in Tables 2, 3, and 4. To obtain statistical data involving SCC-DFTB, three compounds are discarded. F₂ is removed from the training set, as SCC-DFTB yields an unrealistic F-F bond length of 1.112 Å, to be compared with the B3LYP/6-31G* value of 1.404 Å. The OH⁺ and NH₂⁺ cations are not considered either since the data available correspond to their triplet state, which cannot be handled by the tight-binding formalism at the basis of SCC-DFTB.

4. Equilibrium Geometries

4.1. Assessment against B3LYP/6-31G* Structures. The reliability of different levels of theory for the determination of equilibrium geometries is extensively documented in the literature.^{91,92} B3LYP/6-31G* geometries are known to be

Table 4. Same Data as in Table 2 for Anions^a

compound	exptl.	G3MP2B3	PM3	DFT
CN ⁻ (cyanide)	17.7 b	-3.1	9.8	-5.3
C ₆ H ₅ ⁻ (phenyl)	54.7 b	-2.9	-2.9	-0.2
C ₅ H ₅ ⁻ (cyclopentadienyl)	21.3 a	-2.6	-5.3	2.4
HCO ₂ ⁻ (formate)	-110.9 a	-1.7	0.0	-3.8
H ₃ C-CH ₂ ⁻	35.1 b	-1.4	-3.3	0.7
C ₆ H ₅ CO ₂ ⁻	-97.3 b	-1.2	7.4	0.7
C ₆ H ₅ O ⁻	-39.4 b	-1.2	-4.5	-0.5
H ₃ C-N-CH ₃ ⁻	26.1 b	-1.0	-18.2	-2.2
C ₄ H ₄ N ⁻	18.9 b	-1.0	-7.4	1.4
C ₅ H ₅ ⁻	19.6 b	-0.7	-3.6	4.1
HCC ⁻	65.5 b	-0.5	10.8	0.5
C ₆ H ₅ O ⁻ (phenoxy)	-40.5 a	0.0	-3.6	-0.5
CHO ⁻	1.9 b	0.0	-9.3	0.0
NH ₂ ⁻	27.0 b	0.0	11.2	3.1
OH ⁻ (hydroxyde)	-33.2 a	0.0	15.5	2.2
CH ₂ CN ⁻	25.1 b	0.2	3.3	-3.1
CH ₃ NH ⁻	32.0 b	0.2	-10.3	0.5
C ₂ H ₆ N ⁻ (dimethyl nitrogen)	24.7 a	0.2	-16.7	-0.7
H ₂ CCH ⁻	52.8 b	0.5	8.8	2.6
C ₂ H ₃ O ₂ ⁻ (acetate)	-122.5 a	0.7	2.9	-0.5
NO ₂ ⁻	-45.2 b	1.0	2.2	-1.9
CH ₂ NO ₂ ⁻	-27.2 b	1.2	-16.0	-2.2
HOO ⁻	-22.5 b	1.4	-1.4	-0.7
CH ₄ N ⁻ (methylamine)	30.5 a	1.7	-8.6	0.5
C ₅ H ₁₁ ⁻ (neopentyl)	3.2 a	1.9	7.9	9.6
NO ₃ ⁻ (nitrate)	-74.7 a	2.2	-18.4	-2.6
CH ₃ ⁻	33.2 b	2.4	18.2	4.8
C ₂ H ₅ O ⁻ (ethoxy)	-47.5 a	4.5	2.6	2.9
CH ₃ O ⁻ (methoxy)	-36.0 a	4.5	-1.9	3.6
H ⁻ (hydrure)	33.2 b	6.2	58.6	4.8

^a Sources: *a* = ref 15, *b* = Supporting Information from ref 94.

Table 5. Summary of Root Mean Square Deviations of Bond Lengths and Angles Calculated Using Efficient Methods (AM1 and SCC-DFTB) from Corresponding Values Calculated at the B3LYP/6-31G* Level

	lengths (Å)		angles (deg)	
	AM1	SCC-DFTB	AM1	SCC-DFTB
neutrals	0.015	0.025	1.327	2.103
cations	0.043	0.046	7.855	5.720
anions	0.016	0.039	1.606	3.422

quite accurate. In fact, the composite methods G3B3 and G3MP2B3 rely on B3LYP/6-31G* structures.⁴⁴ However, for large ions or when a large number of structures is to be considered, more efficient methods are of interest. Presently available data suggest SCC-DFTB as the method of choice for fast optimization of molecular geometries, except in the presence of NO bonds for which huge errors are noted.^{93,94} However, this conclusion emerges from investigations mainly focused on common CHNO neutrals. Therefore, further assessment of geometries obtained with such methods is of interest, especially for ions.

Experimental gas-phase geometries are clearly not available for most species considered in this work. Therefore, the quality of SCC-DFTB and NDDO structures is assessed against B3LYP/6-31G* geometries. Among present NDDO methods, the best agreement with B3LYP/6-31G* geometries is obtained for AM1 structures. Root mean square deviations from reference B3LYP/6-31G* data for bond lengths and angles are reported in Table 5. Deviations from B3LYP/6-31G* prove systematically larger for ions than for neutral species, and larger for cations than for anions. This latter

Table 6. RMS Deviations from the Experiment of P10-B3LYP/6-311++G** Enthalpies for Neutrals (N), Cations (C), Anions (A), and the Whole Set of Ions (AI)^a

geometry	N	C	A	AI
B3LYP/6-31G*	2.1 (-4.8/+6.0)	5.7	3.3	4.8 (-21.5/+13.4)
AM1	2.3 (-4.8/+7.4)	6.4	5.3	6.0 (-10.0/+25.3)
SCC-DFTB	3.0 (-6.7/+6.0)	5.0	4.0	4.6 (-9.8/+15.1)

^a In addition, minimum and maximum deviations are given in parentheses for neutrals and ions. All data are in kcal/mol. The performance of SCC-DFTB for cations is overestimated by the fact that cations in triplet states are not considered at this level.

finding might be unexpected, since the minimal valence basis set used at the AM1 level might be expected to be a more significant problem for anions than for cations, owing to the more diffuse character of anionic electron clouds. However, it is understandable if the limitations of AM1 for ions arise primarily as a result of the simplifications regarding the Hamiltonian matrix, rather than as a consequence of the lack of flexibility of the basis set.

SCC-DFTB appears significantly less reliable than AM1 for anions and neutral species, with RMSD values for bond lengths and angles about twice as large. This result is unexpected since AM1 is parametrized primarily against data for neutral compounds. Moreover, it contrasts with previous investigations focused on CHNO compounds.^{93,94} However, a detailed examination of present data reveals that the relatively poor performance of SCC-DFTB is due to fluorinated compounds. In particular, C–F bond lengths in F₃C–NF₂ are overestimated by as much as 0.177 Å. Moreover, while N–O bonds are known to be poorly described by SCC-DFTB,^{93,94} it is especially true for the O₂N–OF bond length which is too long by 0.131 Å.

4.2. Influence on Energies and Enthalpies. Total energy increases by up to 6 kcal/mol are observed on substituting B3LYP/6-31G* geometries with either SCC-DFTB or AM1 structures, with many increases in the range 2.5–5 kcal/mol. This is consistent with the root-mean-square increase of about 5 kcal/mol observed previously on substituting DFT geometries with molecular mechanics structures.⁴² Because such energy variations have the same magnitude as typical deviations from experimental values of $\Delta_f H^\circ$ derived from present AE schemes, as detailed below, applying such schemes to approximate geometries would lead to a dramatic loss of accuracy.

However, it is interesting to note that this is not the case provided that the AEs are specifically optimized for these more approximate geometries. Indeed, Table 6 reports the performance of the P10 procedure applied to B3LYP/6-311++G(2df,2p) energies calculated on geometries optimized at lower levels of theory, using AE specifically optimized for the corresponding structures. Only a moderate increase of the RMSD is observed on going from B3LYP/6-31G* structures to either AM1 or SCC-DFTB structures. This shows that reoptimizing the AE parameters is an efficient way to make up for systematic errors affecting the underlying geometries. As a result, the use of single-point calculations on AM1 or SCC-DFTB geometries is reasonable, provided that specific AEs are used. On the other hand, as discussed in the sequel, the best results for ions, and more

Table 7. RMS Deviations from Experiment of B3LYP Enthalpies Calculated Using the P3, P5, P8, and P10 Correction Schemes, with the 6-311++G(2df,2p) and 6-31++G** Bases^a

	N	C	A	AI
6-311++G(2df,2p) basis set				
P3	2.2 (-5.5/+7.4)	6.2	3.6	5.3 (-22.0/+14.8)
P5	2.6 (-4.3/+6.2)	6.3	3.2	5.2 (-22.0/+15.1)
P8	2.3 (-5.3/+5.7)	6.0	3.2	5.0 (-22.2/+13.4)
P10	2.2 (-4.5/+5.7)	5.9	3.2	4.9 (-22.2/+13.4)
6-31++G** basis set				
P3	2.9 (-4.1/+8.8)	6.6	3.7	5.6 (-22.5/+14.3)
P5	2.8 (-4.8/+7.9)	6.7	3.5	5.6 (-22.5/+14.3)
P8	2.7 (-5.5/+6.7)	6.5	3.4	5.4 (-22.7/+13.6)
P10	2.3 (-4.5/+5.5)	6.1	3.4	5.2 (-22.7/+12.7)

^a Minimum and maximum deviations for neutrals and ions are also given in parentheses. All values are reported in kcal/mol.

specifically for anions, are obtained with relatively flexible bases including diffuse functions. Therefore, the use of B3LYP/6-31G* geometries is no dramatic overhead. Unless mentioned otherwise, the following results refer to such structures.

5. Comparison of Present Correction Schemes

While many different correction schemes are used in the literature, systematic comparisons of their relative merits are still lacking. Some results suggest that more flexible schemes yield more reliable $\Delta_f H^\circ$ values.²⁹ However, with a lack of cross-validation or application of the models to external test sets, it is not possible to determine whether the improvement observed on increasing the flexibility of the empirical correction scheme reflects a true enhancement of the predictive power of the method.

For all theoretical levels employed in the present work to compute total energies, the procedures P3, P5, P8, and P10 have been applied to convert total energies into formation enthalpies. Whatever the specific procedure used, quite similar results are obtained. Nevertheless, the RMSD values reported in Table 7 indicate a small but mostly systematic improvement with the number of adjustable parameters, with the P10 and P8 procedures yielding the smallest RMSD values. On the other hand, no improvement is noted concerning the minimum and maximum deviations from experimental values. Similar observations can be made for other basis sets and functionals. The only exceptions concern the functionals for which the assumption at the basis of the P3 correction scheme breaks down, such as PBE1PBE and to some extent PBE (c.f. section 8). In such cases, P3 naturally leads to very poor predictions. The fact that going from P5 to P10 provides only a few improvements indicates that the P8/P10 definitions for the AEs account only for a small fraction of their environment dependence. Alternatively, one might consider introducing even more specific AEs, or group equivalents, which appear to provide some improvement for neutrals.^{41,42} However, their determination requires a training set larger than the present one. Moreover, this approach does not address the fundamental limitations of equivalents associated with their local character.

Table 8. RMS Deviations (kcal/mol) from Experimental Results of B3LYP Enthalpies Calculated Using the P5 Procedure and a Variety of Basis Sets, for Neutrals Compounds of the Training Set (N), Cations (C), Anions (A) and All Ions (AI) of the Test Set^a

	N	C	A	AI	CPU
6-31G*	4.9	7.7	26.1	17.9	273
6-31G**	3.9	7.5	25.8	17.7	338
6-311G*	4.0	6.5	16.7	11.9	477
6-31+G*	3.6	7.3	9.7	8.4	731
6-31+G**	2.9	6.7	9.6	8.1	861
6-31++G**	2.8	6.7	3.5	5.6	1022
6-311+G*	3.3	6.7	5.3	6.1	1211
6-311+G**	2.5	6.1	4.9	5.6	1418
6-311++G**	2.5	6.0	3.4	5.1	1663
6-311++G(2df,2p)	2.6	6.3	3.2	5.2	7803

^a The last column provides a rough indication of the relative cost of the different bases as implemented in Gaussian. Each CPU number represents the relative CPU time of a single point calculation for a typically energetic salt, namely, a bicyclic azolium with empirical formula $C_7H_9N_6O_2$ and no symmetry.⁹⁵ An AM1 calculation using Gaussian corresponds roughly to CPU = 1.

6. Influence of the Basis Set

It is well-known that relatively extended basis sets including diffuse functions are required for an accurate description of the electronic structure and properties of anions. The present work shows that this conclusion remains valid for $\Delta_f H^\circ$ after application of the present correction schemes. In other words, AEs are not efficient to make up for basis set deficiencies. This observation is illustrated in Table 8 by the results obtained with the B3LYP functional associated with the P10 procedure.

For neutrals, cations, and anions, $\Delta_f H^\circ$ predictions steadily improve as more flexible bases are used. For neutrals and cations, the basis set limit appears to be obtained using 6-311+G*, with 6-31+G** already providing quite good results. For anions, diffuse functions on hydrogen atoms, usually considered to play a marginal role,⁹² prove necessary to obtain RMSD values below 3.5 kcal/mol. Increasing the flexibility of the basis beyond 6-311++G** still yields some improvement for anions. However, it is not very significant, since the RMSD decreases only by 0.2 kcal/mol if 6-311++G(2df,2p) is used instead. Therefore, one might prefer the former basis which is about 5 times more efficient.

Not surprisingly, a flexible basis set including diffuse functions is especially mandatory for anions. RMSD values > 15 kcal/mol are obtained otherwise. With the most flexible bases, comparable RMSDs (2.6–3.2 kcal/mol) are obtained for neutrals and anions, while the corresponding value (6.3 kcal/mol) for cations is significantly larger, as discussed in further detail in section 9.

While diffuse functions are especially important for anions, they prove significant for neutral compounds as well, with the RMSD decreasing from 3.9 to 2.9 kcal/mol on going from 6-31G** to 6-31+G**. Not surprisingly, their role is not significant for cations which exhibit more compact electron clouds.

Focusing on neutrals, the deviations from experimental results increase 2-fold on going from the most flexible to the smallest basis. The RMSD of 3.9 kcal/mol obtained using

Table 9. RMS Deviations (kcal/mol) from Experimental Results of Enthalpies Calculated Using the P10 Procedure and the 6-31++G** Basis Set^a

functional	N	C	A	AI
Xalpha	3.0	22.7	20.6	21.8
SVWN	3.0	15.6	9.6	13.4
BLYP	3.4	8.6	4.8	7.3
BP86	3.3	7.2	3.4	6.0
PBEPBE	3.3	7.5	4.8	6.5
mPWPW91	3.3	7.3	4.1	6.2
PW91PW91	3.2	7.3	3.8	6.1
HCTH	3.3	7.0	3.8	5.9
HF	4.8	26.1	19.2	24.2
TPSS	3.7	7.6	5.5	6.8
VSXC	2.7	7.1	5.5	6.5
BHandH	3.0	9.8	9.8	9.8
BHandHLYP	2.6	6.9	7.4	7.1
B1LYP	2.2	7.0	5.5	6.4
B3P86	2.2	13.9	10.7	12.6
PBE1PBE	2.1	6.0	4.8	5.5
B3LYP	2.3	6.1	3.4	5.2
B3PW91	2.2	5.8	3.9	5.1
B98	2.2	6.1	4.3	5.4
B1B95	2.1	6.3	6.4	6.3
BMK	2.8	6.4	6.6	6.5

^a The functionals are listed from top to bottom according to the underlying approximation, starting with LDA and followed by GGA, HF, M-GGA, H-GGA, and HM-GGA.

the popular 6-31G** basis is significantly larger than the value of 2.5 kcal/mol obtained with the most flexible bases. On the other hand, a numerical basis set such as DN** proves superior to a Gaussian basis of similar flexibility, such as 6-31G**, in line with previous results.²⁰ This finding is not surprising as a numerical basis better spans the space of the s and p atomic orbitals.

Finally, 6-31++G** appears to provide a reasonable trade-off between efficiency and accuracy if anions are to be considered, while 6-31+G** should be sufficient otherwise. Indeed, it yields RMSD values within 0.5 kcal/mol of those obtained using 6-311++G(2df,2p) which is almost 8 times more costly. This good performance is observed for neutral compounds, for cations, and for anions as well.

7. Influence of the Functional

A comparison of the performance of various functionals for $\Delta_f H^\circ$ prediction is provided in Table 9, using the P10 procedure and the 6-31++G** basis set. The best results are obtained using popular H-GGA functionals, especially B3LYP and B3PW91, with RMSD < 5.3 kcal/mol for ions and < 2.4 kcal/mol for neutrals. Although PBE1PBE results are better for neutrals, they appear less reliable for ions. Other H-GGA functionals, especially B3P86, BHandH, and to a lesser extent BHandHLYP do not perform so well. The relatively poor results obtained using BHandHLYP may be attributed to the overestimated contribution of HF exchange in this inappropriately constructed functional. Similarly, the results obtained using the more complicated HM-GGA approach are somewhat less satisfactory. All in all, among the DFT functionals considered in this paper, B3LYP, B3PW91, B98, and PBE1PBE emerge as the most reliable for $\Delta_f H^\circ$ predictions on ionic systems.

Table 10. Atom Equivalents for the P5 Procedure with the 6-31++G** Basis Set (kcal/mol)^a

functional	H	C	N	O	F	R ²
Xalpha	-30.6	-274.4	-332.1	-390.9	-462.2	1.00
SVWN	2.1	-98.1	-133.2	-171.2	-225.7	0.98
BLYP	-2.0	33.9	42.7	50.9	51.7	1.00
BP86	1.4	41.4	50.0	57.1	55.0	1.00
PBEPBE	-1.8	13.1	12.9	10.8	-1.9	0.73
mPWPW91	0.4	39.4	46.5	52.7	50.0	1.00
PW91PW91	-0.4	33.7	39.2	44.1	39.6	1.00
HCTH	5.7	35.8	39.3	41.7	37.8	0.99
HF	-10.2	-107.5	-145.0	-183.3	-203.9	0.98
TPSSTPSS	1.6	46.5	53.3	59.5	58.9	1.00
VSXC	1.9	52.9	57.6	64.7	67.8	0.99
BHandH	-11.0	-134.5	-176.1	-219.5	-263.6	0.99
BHandHLYP	1.1	27.5	26.1	25.0	29.3	0.84
B1LYP	-0.4	30.4	34.0	37.3	39.9	0.99
B3P86	14.6	108.4	124.4	139.6	150.9	1.00
PBE1PBE	0.2	14.1	10.0	4.4	-4.8	0.02
B3LYP	2.7	41.5	46.7	51.4	53.8	0.99
B3PW91	2.7	31.8	33.2	33.4	30.4	0.95
B98	2.1	31.5	33.6	34.5	33.2	0.96
B1B95	-1.1	32.8	36.1	38.3	40.1	0.98
BMK	0.2	25.8	29.7	30.4	32.2	0.98

^a The functionals are listed from top to bottom according to the underlying approximation, starting with LDA and followed by GGA, HF, M-GGA, H-GGA, and HM-GGA. The last column reports the squared correlation coefficients between AE values and atomic numbers for CNOF atoms.

Turning our attention to more efficient models, the best results are obtained within GGA. BP86 and HCTH, in particular, yield quite satisfactory results, comparable to those obtained with the BMK functional, which is the most costly considered in this paper (about 3 times more time-consuming than SVWN using Gaussian). In fact, GGA appears as valuable as HM-GGA as far as ions are concerned, although BLYP proves somewhat less reliable than other GGA functionals. This result contrasts with what is observed for neutral compounds from the training set, where HM-GGA proves more reliable than GGA, especially if the B1B95 functional is used. Besides such general trends, significant conclusions regarding the relative performance of various functionals within the same family can hardly be drawn from present data in view of the small differences between corresponding RMSD data.

8. Atom Equivalents

To discuss the main features of atom equivalents as defined in the present paper, values obtained for the P5 procedure, using the 6-31++G** basis set and various functionals, are listed in Table 10. In addition, the variation of environmentally dependent atom equivalents for the P10 procedure, using the same basis set, is illustrated in Figure 1 for the functionals considered in this paper. All AEs derived in this work are statistically well-defined. Indeed, the associated standard deviations derived from a singular value decomposition are systematically $\ll 1\%$ of their actual values.

As stated in section 2.1, the explicit evaluation of all contributions to $\Delta_f H^\circ$ allows a straightforward interpretation of these values in terms of additive corrections to E_0 . Their magnitude is much smaller than additive contributions to $\Delta_f H^\circ - E_0$ often used in practical schemes.⁴² On the other hand, it is much larger than additive corrections to theoretical

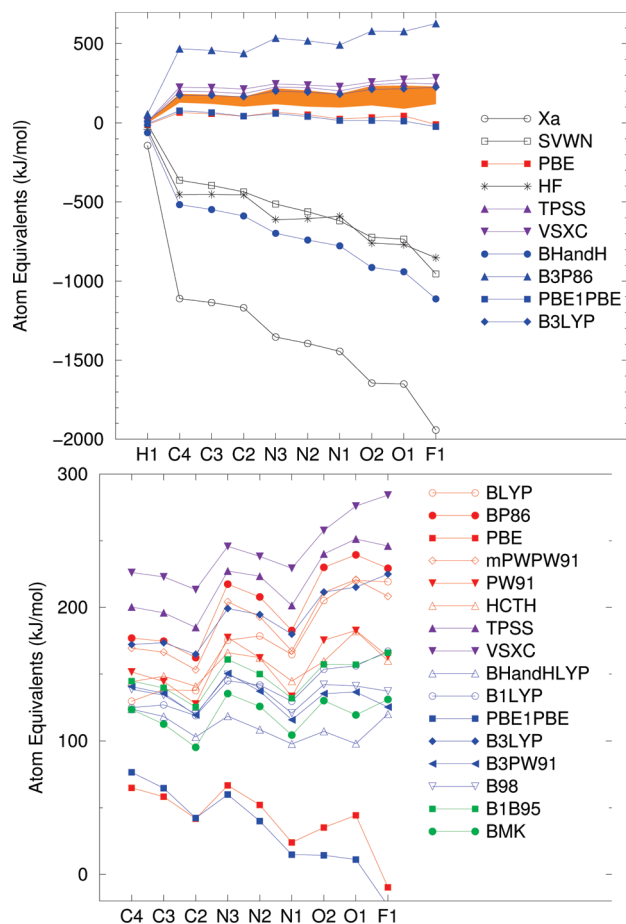


Figure 1. Dependence of P10 atom equivalents for use with 6-31++G** on the energy functional. Color code for the functionals: white = LDA, red = GGA, indigo = M-GGA, blue = H-GGA, and green = HM-GGA. In the upper graph, Xa stands for the X α functional, and the shaded area is where data symbols for most GGA, H-GGA, and HM-GGA functionals are to be found, as detailed in the lower graph.

formation enthalpies derived from molecular reference states.⁴⁰ This is understandable as errors on total energies partially cancel out for such reference states.

While the reliability of calculated enthalpies depends mostly on the basis set, AE values exhibit a more significant dependence on the functional. Focusing on 6-31++G** data, they increase according to the following order: X α \ll BHandH < HF \approx SVWN \ll PBE \leq PBE1PBE < most DFT functionals < TPSS < VSXC \ll B3P86, with negative values for the former four methods. This significant dependence indicates that AE corrections are suitable to make up for systematic errors associated with the functional but not so efficient to correct for basis set effects.

Considering all parameters as a whole, the upper plot in Figure 1 clearly shows that the main difference is between the AE for hydrogen and those for heavier elements. With regard to the latter, their different values turn out to depend on both the number of electrons and the coordination number of the atom. For some models requiring relatively large corrections to E_0 , such as X α , SVWN, or B3P86, the atomic number have the most significant influence on AE values. However, for most functionals, the role of the coordination number is equally significant, as is clear from the bottom

graph in Figure 1. This confirms the actual significance of the slight improvement observed on going from the P5 to the P10 correction scheme.

An interesting feature is the linear correlation observed between AE values and atomic numbers, at the basis of the present P3 procedure. Considering AEs derived for P5 and listed in Table 10, this correlation is usually very high, with squared correlation coefficients $R^2 \geq 0.99$ in most cases. This explains the similar performances of P3 and P5 procedures for most functionals. However, no such correlation is observed for PBE1PBE, while it is relatively poor ($R^2 = 0.71$) for PBE.

Considering the influence of the atomic environment, it is clear from Figure 1 that, in most cases, AE values increase with the atom coordination number. The reverse dependence is observed only in few cases, for instance, for oxygen using HCTH or VSXC and for nitrogen at the HF level. This indicates that for most DFT functionals, namely those associated with positive equivalents, errors on total energies are less significant for atoms with lower coordination numbers, while the opposite is true for $X\alpha$, SVWN, and BHandH, which require negative equivalents.

At the HF level, all atom equivalents are negative, as expected in view of the variational character of the Hartree–Fock theory, which implies overestimated E_0 values. Interestingly, notwithstanding the role of coordination numbers, the corrections to SVWN energies are quite similar. This might appear paradoxical in view of the very different behaviors of HF and LDA approximations, for instance, the fact that HF tends to underestimate bond energies, while the opposite is true for LDA functionals.⁹¹ However, it must be kept in mind that present corrections apply to total energies, not to binding energies. Therefore, the similar magnitudes of HF and LDA equivalents is related to the fact that both theories yield similar errors on molecular total energies, while total energies of isolated atoms are more severely overestimated at the LDA level.

Compared with HF or SVWN functionals, the more approximate $X\alpha$ method requires corrections roughly twice as negative, indicating that E_0 values are even more overestimated. On the other hand, the fact that $X\alpha$ bond dissociation energies are not significantly worse than SVWN values⁹¹ indicates that the overestimation of E_0 with respect to SVWN is observed for isolated atoms as well. The last functional for which significantly negative atom equivalents are obtained is BHandH. This functional is unique among advanced DFT functionals with regard to its overestimation of total energies. All other functionals yield essentially positive AEs, i.e., underestimated E_0 values. PBE1PBE and, to a lesser extent, PBE emerge as the functionals requiring the least significant corrections. For energies calculated using the 6-31++G** basis, this is clear considering the AEs listed in Table 10 for P5 corrections and those shown in Figure 1 for P10 corrections. Of course, since these functionals underestimate E_0 , increasing the size of the basis set decreases E_0 and therefore calls for larger corrections.

Because the magnitude of AEs reflects errors of the functionals on calculated *total energies*, while they are more often assessed on the basis of a comparison of *energy*

Table 11. Summary of RMS Deviations (kcal/mol) between Theoretical and Experimental Enthalpies for Sets Made of Neutral Compounds (N), Cations (C), Anions (A), and All Ions (AI)

	N	C	A	AI
G3MP2B3	1.9	4.4	2.4	3.7
P10–B3LYP/6-311++G**	2.1	5.7	3.3	4.8
P5–B3LYP/6-311++G**	2.5	6.0	3.4	5.1
P10–B3LYP/6-31++G**	2.3	6.1	3.4	5.2
P5–B3LYP/6-31++G**	2.8	6.7	3.5	5.6
AM1	11.0	9.6	15.2	12.3
PM3	7.0	13.3	14.2	13.7
RM1	12.3	25.6	26.9	26.2
P5-SCC-DFTB	14.5	14.9	28.8	16.2
P10-SCC-DFTB	11.5	11.8	31.1	22.3

differences with experimental data, the present study provides some new insight into the relative performances of different functionals. Present results should also be useful in view of extending available AE schemes to new elements while minimizing the number of adjustable parameters.

9. Calculated Enthalpies

This final section examines in more detail the enthalpies calculated using the various procedures considered in this paper. For this purpose, PM3, G3MP2B3, and P5-B3LYP/6-31++G** enthalpies are compared to experimental values in Tables 2, 3, and 4 for neutrals, cations, and anions, respectively. An overview of the relative performance of some of the most interesting AE schemes presently introduced is provided in Table 11, where they are compared to G3MP2B3 and NDDO methods.

9.1. G3MP2B3 Enthalpies. In order not to spoil present AE values with the use of spurious data, only compounds with experimental $\Delta_f H^\circ$ values in reasonable agreement with G3MP2B3 values are included in the training set. As a result, all G3MP2B3 values reported in Table 2 are within 7 kcal/mol from experimental ones, in line with the usual performance of the G3MP2B3 method.⁹⁷ In fact, the largest deviations arise for compounds for which the NIST values are reported with significant error bars (up to 2.5 kcal/mol for instance for F_3C-OF).

While an accurate description of the electronic structure of anions might be expected to prove more challenging owing to its diffuse character, $\Delta_f H^\circ$ values for anions are quite satisfactory, with RMSD = 2.4 kcal/mol and all deviations between -6 kcal/mol (for $HCOO^-$) and $+6.3$ kcal/mol (for H^-), as shown in Table 4. However, somewhat larger deviations are observed for cations, as reported in Table 3. In particular, the experimental value found for the cyclobutyl cation is 18 kcal/mol above experimental results. This disagreement might stem from the fact that this cation can easily undergo transition to isomers about 9 kcal/mol lower in energy.⁹⁹ In fact, the RMSD for cations, reported as 4.4 kcal/mol in Table 11, drops to 3.5 kcal/mol if the cyclobutyl cation is not considered. This value remains almost 50% larger than the corresponding value for anions. The larger RMSD obtained for cations cannot be explained by the occurrence of triplet species (OH^+ and NH_2^+) in the cation data set, as deviations from the experiment for these open

Table 12. Atom Equivalents for the P10 Procedure (kcal/mol)

model for E_0	H	C4	C3	C2	N3	N2	N1	O2	O1	F
For Use with B3LYP/6-31G* Geometries										
B3LYP/6-31+G**	2.7	41.2	41.5	39.4	47.6	46.5	43.0	50.5	51.4	53.8
B3LYP/6-31++G**	2.7	41.2	41.5	39.4	47.6	46.5	43.0	50.5	51.4	53.8
B3LYP/6-311++G**	3.3	44.3	45.3	44.7	55.3	54.1	52.2	63.6	65.0	71.3
B3LYP/6-311++G(2df,2p)	3.3	45.8	47.0	46.3	57.6	56.2	54.1	66.5	67.8	74.5
For Use with AM1 Geometries										
B3LYP/6-311++G(2df,2p)	3.0	45.8	47.1	46.5	56.1	55.0	54.1	64.1	67.9	73.6
B3LYP/6-31+G**	2.6	41.5	41.6	39.5	47.9	46.3	43.0	50.5	51.1	53.6
For Use with SCC-DFTB Geometries:										
B3LYP/6-311++G(2df,2p)	2.6	47.3	47.5	45.1	58.4	55.0	52.5	66.6	66.6	72.6
B3LYP/6-31+G**	1.8	43.2	42.2	38.0	48.9	45.3	39.9	51.0	49.7	51.7

shell systems are not especially large. According to the data in Tables 3 and 4, it arises because of relatively large deviations (>5 kcal/mol) for some cyclic species (cyclopropyl, cyclopentyl, cyclohexyl). All in all, the RMSD increases by ca. 60% on going from neutrals to ions. This increase might stem to some extent from larger uncertainties associated with experimental data for gas-phase ions.

For neutrals, the RMSD between G3MP2B3 and experimental enthalpies can be further decreased by the application of additive corrections. In the present case, bond equivalents⁹⁷ do not perform significantly better than P10 corrections, in view of the respective RMSD values of 1.5 and 1.6 kcal/mol. In fact, this specific case of G3MP2B3 calculations for neutrals is the only one where BE corrections are found to provide better results than AE corrections. However, no correction scheme was found to improve G3MP2B3 enthalpies for ions. Therefore, one should consider raw G3MP2B3 values rather than bond-corrected values for studies involving ionic systems.

9.2. Semiempirical Enthalpies. PM3 and RM1 were developed in order to overcome the limitations of AM1 to predict $\Delta_f H^\circ$. In fact, Table 11 clearly shows that none of these semiempirical methods is suitable for ions. In fact, PM3 and RM1 predictions for ions are even worse than AM1 values. This is especially true for RM1, probably as a result of its more empirical character. Furthermore, although this recent method is reported to yield a significant improvement over previous NDDO schemes with regard to the prediction of formation enthalpies for neutral organic compounds,¹⁵ it proves even worse for neutral compounds in the present training set. This may be explained by the fact that the large data sets employed to fit RM1 and earlier NDDO methods are not representative of the variety of bonding patterns spanned by the present training set. While none of these methods are parametrized for ions, the fact that deviations from experimental results for ions are twice as large with RM1 than with AM1 or PM3 confirms the idea that the superiority of RM1 for standard organic compounds is obtained at the expense of its reliability for less commonly encountered structures. PM3 appears to provide a reasonable trade-off between reliability for common structures and applicability to less common moieties.

All in all, present NDDO results for ions are not significantly better than those obtained using obsolete schemes such as MINDO/3 or MNDO.⁹⁸ Indeed, accordingly to the RMSD data in Table 11, typical errors are about 12

kcal/mol, or somewhat lower for neutrals using PM3. They are at best about 4–6 times larger than G3MP2B3 errors. This confirms the potential interest of procedures more reliable than available NDDO approaches for ions, while at the same time being more efficient than G3MP2B3 for complex systems.

9.3. Enthalpies Derived from Present AE Schemes. The combination of DFT energies with AEs provides such procedures. For instance, it is clear from Tables 2, 3, and 4 that P10 equivalents applied to B3LYP/6-31++G** energies provide $\Delta_f H^\circ$ values much more reliable than PM3 values. The improvement is dramatic for small molecules including HN=HN, N₂, and F₂; some medium-size molecules such as isoxazole; and small ions such as NO₂⁺, OH⁻ and H⁻, including some triplet species, OH⁺ and NH₂⁺. With respect to G3MP2B3, such a DFT/AE approach yields errors typically 20% larger for neutrals and 40% larger for ions, as clear from Table 11. Using 6-311++G** instead of 6-31++G**, the corresponding increases are respectively 10% and 30%. Therefore, notwithstanding the possibility of larger uncertainties associated with data for ions, their enthalpies prove more difficult to predict than values for neutrals with DFT/AE procedures.

The lower accuracy of present DFT/AE approaches compared with G3MP2B3 is consistent with recent findings based on group equivalents optimized for nitro compounds.^{18,41} Applied to present ions, the latter yield respectively 6.8 and 5.2 kcal/mol for RMSD data associated with cations and anions. These values are larger than present ones reported in Table 11, as expected from their specialization toward energetic materials. However, these data confirm present conclusions regarding application of DFT/AE to ions, such as the largest deviations observed for cations and the clear superiority of such approaches over NDDO schemes.

Many combinations of the functional/basis set/correction scheme yield very similar results, with the best ones obtained using popular H-GGA functionals: B3LYP, B3PW91, B98, and PBE1PBE, followed by GGA functionals BP86 and HCTH. Whatever the correction scheme employed, SCC-DFTB yields especially poor results, especially for anions. This is probably related to the localized basis set specific to this method.⁴⁸ For practical purposes, application of the P10 equivalents listed in Table 12 for B3LYP/6-31++G** or B3LYP/6-311++G** energies appears to be a valuable procedure. P10-B3LYP/6-31++G** is now the default procedure for ions in the MATEO program.²²

10. Conclusion

The present work is the most comprehensive investigation to date of the relative performance of various approaches to estimate the formation enthalpies of ions. In particular, it provides a clear overview of the accuracy to be expected from AE schemes, and new insight into their relative merits.

The most reliable DFT/AE procedures yield a root-mean-square deviation (RMSD) from experimental values < 2.5 kcal/mol for neutrals, but close to 5 kcal/mol for a database of 73 ions. This loss of accuracy is mainly attributed to the fact that present local corrections cannot capture the electron delocalization in ions, associated with the fact that their electron cloud typically involves several canonical structures. Allowing a dependence of AE parameters on the coordination number of the atoms provides a small but systematic improvement. Therefore, such coordination-dependent parameters should preferably be used, unless nonequilibrium structures such as transition states are to be considered.

On the other hand, approximate geometries derived from relatively low theoretical levels, such as AM1 or SCC-DFTB, may be used with no significant loss of accuracy with respect to calculation on B3LYP/6-31G* geometries provided that the AEs employed are specifically optimized for these more approximate geometries. Although AE corrections prove quite efficient, and have been recently shown to make up for errors associated with the use of small basis sets for many systems,^{19,20} present results indicate that this is not really the case for anions, with the associated RMSD steadily increasing as less flexible bases are considered. This is bad news for practical applications. Indeed, unless more sophisticated correction schemes are introduced, this result implies that routine calculations of formation enthalpies for ionic liquids or energetic salts require relatively costly basis sets including diffuse functions in order to get the most from DFT.

Nevertheless, although more efficient procedures are desirable for routine calculations, present DFT/AE procedures fill the gap between costly composite methods and unreliable semiempirical schemes. In particular, the use of B3LYP/6-31++G** energies provides a good trade-off between reliability and cost for anions and cations. In view of predicting the performance of energetic salts, such procedures appear especially suitable. Indeed, as long as the reliability of such predictions depends on the uncertainties associated with the evaluation of lattice energies, using more sophisticated procedures to compute gas phase enthalpies will not necessarily lead to significant improvement.

Acknowledgment. The authors thank Dr. M. Elstner for the SCC-DFTB code used in this work. The authors also acknowledge fruitful comments from anonymous reviewers.

Supporting Information Available: Molecular files containing B3LYP/6-31G* optimized geometries in .xyz format. Theoretical G3MP2B3 thermochemical data are included as comments in these files (in atomic units). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Freemantle, M.; Welton, T.; Rogers, R. D. *An Introduction to Ionic Liquids*; Royal Society of Chemistry: London, 2009.
- (2) Guo, Y.; Tao, G.-H.; Joo, Y.-H.; Wang, R.; Twamley, B.; Parrish, D. A.; Shreeve, J. M. *Energy Fuels* **2009**, *23*, 4567–4574.
- (3) Klapoetke, T. M.; Sabate, C. M. Z. *Anorg. Allg. Chem.* **2009**, *635*, 1812–1822.
- (4) Irikura, K. K.; Frurip, D. J. *Computational thermochemistry: prediction and estimation of molecular thermodynamics*; American Chemical Society: Washington, DC, 1998.
- (5) Allinger, N. L.; Yan, L. *J. Am. Chem. Soc.* **1993**, *115*, 11918–11925.
- (6) Mathieu, D.; Simonetti, P. *Mol. Eng.* **1999**, *8*, 121–134.
- (7) Sukhachev, D. V.; Pivina, T. S. *Propellants Explos. Pyrotech.* **1994**, *19*, 159–164.
- (8) Marino, D. J. G.; Peruzzo, P. J.; Krenkel, G.; Castro, E. A. *Chem. Phys. Lett.* **2003**, *369*, 325–334.
- (9) Gutowski, K.; Holbrey, J.; Rogers, R.; Dixon, D. *J. Phys. Chem. B* **2005**, *109*, 23196–23208.
- (10) Gutowski, K.; Rogers, R.; Dixon, D. *J. Phys. Chem. B* **2006**, *110*, 11890–11897.
- (11) Gao, H.; Ye, C.; Piekarski, C. M.; Shreeve, J. M. *J. Phys. Chem. C* **2007**, *111*, 10718–10731.
- (12) Wang, L.; He, Y.-L. *Int. J. Mass Spectrom.* **2008**, *276*, 56–76.
- (13) Curtiss, L.; Raghavachari, K. *Theor. Chem. Acc.* **2002**, *108*, 61–70.
- (14) DeYonker, N. J.; Grimes, T.; Yockel, S.; Dinescu, A.; Mintz, B.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *125*, 104111.
- (15) Rocha, G. B.; Freire, R. O.; Simos, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101–1111.
- (16) Repasky, M.; Chandrasekhar, J.; Jorgensen, W. *J. Comput. Chem.* **2002**, *23*, 1601–1622.
- (17) Beaucamp, S.; Bernand-Mantel, A.; Mathieu, D.; Agafonov, V. *Mol. Phys.* **2004**, *102*, 253–258.
- (18) Byrd, E. F. C.; Rice, B. M. *J. Phys. Chem. A* **2009**, *113*, 345–352.
- (19) Brothers, E. N.; Scuseria, G. E. *J. Chem. Theory Comput.* **2006**, *2*, 1045–1049.
- (20) Delley, B. *J. Phys. Chem. A* **2006**, *110*, 13632–13639.
- (21) Beaucamp, S.; Mathieu, D.; Agafonov, V. *J. Phys. Chem. B* **2005**, *109*, 16469–16473.
- (22) Mathieu, D. *J. Hazard. Mater.* **2010**, *176*, 313–322.
- (23) Friesner, R. A.; Knoll, E. H. *J. Chem. Phys.* **2006**, *125*, 124107.
- (24) Schwabe, T.; Grimme, S. *Acc. Chem. Res.* **2008**, *41*, 569–579.
- (25) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849–1868.
- (26) Brothers, E. N.; Scuseria, G. E. *J. Phys. Chem. A* **2008**, *112*, 13706–13711.
- (27) Cioslowski, J.; Liu, G.; Piskorz, P. *J. Phys. Chem. A* **1998**, *102*, 9890–9900.

- (28) Duchowicz, P.; Castro, E. *Arkivoc* **2001**, 2, 227–241.
- (29) Guthrie, J. J. *Phys. Chem. A* **2001**, 105, 9196–9202.
- (30) Ruzsinszky, A.; Van Alsenoy, C.; Csonka, G. *J. Phys. Chem. A* **2003**, 107, 736–744.
- (31) Ruzsinszky, A.; Csonka, G. *J. Phys. Chem. A* **2003**, 107, 8687–8695.
- (32) Wang, X.; Wong, L.; Hu, L.; Chan, C.; Su, Z.; Chen, G. *J. Phys. Chem. A* **2004**, 108, 8514–8525.
- (33) Long, D. A.; Anderson, J. B. *Chem. Phys. Lett.* **2005**, 402, 524–528.
- (34) Friesner, R. A.; Knoll, E. H.; Cao, Y. *J. Chem. Phys.* **2006**, 125, 124107.
- (35) Yan, G.-K.; Li, J.-J.; Li, B.-R.; Hu, J.; Guo, W.-P. *J. Theor. Comput. Chem.* **2007**, 6, 495–509.
- (36) Goldfeld, D. A.; Bochevarov, A. D.; Friesner, R. A. *J. Chem. Phys.* **2008**, 129, 214105.
- (37) Wu, J.; Xu, X. *J. Chem. Phys.* **2007**, 127, 214105.
- (38) Wu, J.; Xu, X. *J. Comput. Chem.* **2009**, 30, 1424–1444.
- (39) Rice, B.; Pai, S.; Hare, J. *Combust. Flame* **1999**, 118, 445–458.
- (40) Politzer, P.; Ma, Y.; Lane, P.; Concha, M. C. *Int. J. Quantum Chem.* **2005**, 105, 341–347.
- (41) Byrd, E. F. C.; Rice, B. M. *J. Phys. Chem. A* **2006**, 110, 1005–1013.
- (42) Rousseau, E.; Mathieu, D. *J. Comput. Chem.* **2000**, 21, 367–379.
- (43) Ruzsinszky, A.; Kristyan, S.; Margitfalvi, J. L.; Csonka, G. I. *J. Phys. Chem. A* **2003**, 107, 1833–1839.
- (44) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, 110, 7650–7657.
- (45) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, 109, 7764–7776.
- (46) Winget, P.; Clark, T. *J. Comput. Chem.* **2004**, 25, 725–733.
- (47) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; McGraw-Hill: New York, 1989.
- (48) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, 58, 7260–7268.
- (49) Slater, J. C. *The Self-Consistent Field for Molecules and Solids*; McGraw-Hill: New York, 1974; Vol. 4.
- (50) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, 58, 1200–1211.
- (51) Becke, A. D. *Phys. Rev. A* **1988**, 38, 3098–3100.
- (52) Perdew, J. P. *Phys. Rev. B* **1986**, 33, 8822–8824.
- (53) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, 37, 785–789.
- (54) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, 157, 200–206.
- (55) Perdew, J. P. *Unified Theory of Exchange and Correlation Beyond the Local Density Approximation*; Akademie Verlag: Berlin, Germany, 1991.
- (56) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, 46, 6671–6687.
- (57) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, 108, 664–675.
- (58) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, 77, 3865–3868.
- (59) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, 78, 1396–1396.
- (60) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2001**, 114, 5497–503.
- (61) Becke, A. D. *J. Chem. Phys.* **1996**, 104, 1040–1046.
- (62) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, 110, 6158–6159.
- (63) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1997**, 274, 242–250.
- (64) Schmider, H. L.; Becke, A. D. *J. Chem. Phys.* **1998**, 108, 9624–9631.
- (65) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
- (66) Becke, A. D. *J. Chem. Phys.* **1992**, 98, 1372–1377.
- (67) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, 91, 146401.
- (68) Voorhis, T. V.; Scuseria, G. E. *J. Chem. Phys.* **1998**, 109, 400–410.
- (69) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, 121, 3405–3416.
- (70) Vauthier, E.; Blain, M.; Odiot, S.; Barone, V.; Comeau, M.; Fliszar, S. *THEOCHEM* **1995**, 340, 63–70.
- (71) Zope, R. R.; Dunlap, B. I. *J. Chem. Theory Comput.* **2005**, 1, 1193–1200.
- (72) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, 111, 10439–10452.
- (73) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, 120, 215–241.
- (74) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, 107, 3902–3909.
- (75) Stewart, J. J. P. *J. Comput. Chem.* **1989**, 10, 109–220.
- (76) Stewart, J. J. P. *J. Comput. Chem.* **1989**, 10, 221–264.
- (77) mopac7. <http://sourceforge.net/projects/mopac7> (accessed Apr 21, 2010).
- (78) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, 94, 7221–7230.
- (79) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, 106, 1063–1079.

- (80) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374–7383.
- (81) Cioslowski, J.; Schimeczek, M.; Liu, G.; Stoyanov, V. *J. Chem. Phys.* **2000**, *113*, 9377–9389.
- (82) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005**, *123*, 124107.
- (83) Stewart, J. J. P. *J. Comput. Phys.* **1989**, *10*, 221.
- (84) Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495–506.
- (85) Herndon, W. C. *Chem. Phys. Lett.* **1995**, *234*, 82–86.
- (86) Mole, S.; Zhou, X.; Liu, R. *J. Phys. Chem.* **1996**, *100*, 14665–14671.
- (87) Osmont, A.; Catoire, L.; Goekalp, I. *Energy Fuels* **2008**, *22*, 2241–2257.
- (88) Schmitz, L. R.; Motoc, I.; Bender, C.; Labanowski, J. K.; Allinger, N. L. *J. Phys. Org. Chem.* **1992**, *5*, 225–229.
- (89) Schmitz, L.; Chen, K.; Labanowski, J.; Allinger, N. *J. Phys. Org. Chem.* **2001**, *14*, 90–96.
- (90) NIST Chemistry Web Book. <http://webbook.nist.gov/chemistry> (accessed Apr 21, 2010).
- (91) Ziegler, T. *Chem. Rev.* **1991**, *91*, 651–667.
- (92) Foresman, J. B.; Frisch, A. *Exploring Chemistry with Electronic Structure Methods*; Gaussian, Inc.: Pittsburg, PA, 1993.
- (93) Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. A* **2006**, *110*, 13551–13559.
- (94) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751–5755.
- (95) This is 4-Methyl-1-(1-methyl-4-nitroimidazol-2-yl)-1,2,4-triazolium.
- (96) Biehl, H.; Stuhl, F. *J. Chem. Phys.* **1994**, *100*, 141–145.
- (97) Bharthwaj, B.; Melius, C. F. *J. Phys. Chem. A* **2005**, *109*, 1734–1747.
- (98) Halim, H.; Heinrich, N.; Koch, W.; Schmidt, J.; Frenking, G. *J. Comput. Chem.* **1986**, *7*, 93–104.
- (99) Koch, W.; Liu, B.; Defrees, B. J. *J. Am. Chem. Soc.* **1988**, *110*, 7325–7328.

CT100024R

Ligand Entropy in Gas-Phase, Upon Solvation and Protein Complexation. Fast Estimation with Quasi-Newton Hessian

S. Wlodek,* A. G. Skillman, and A. Nicholls

*OpenEye Scientific Software Incorporated, 9 Bisbee Court, Suite D,
Santa Fe, New Mexico 87508*

Received February 17, 2010

Abstract: A method of rapid entropy estimation for small molecules in vacuum, solution, and inside a protein receptor is proposed. We show that the Hessian matrix of second derivatives built by a quasi-Newton optimizer during geometry optimization of a molecule with a classical molecular potential in these three environments can be used to predict vibrational entropies. We also show that a simple analytical solvation model allows for no less accurate entropy estimation of molecules in solution than a physically rigorous but computationally more expensive model based on Poisson's equation. Our work also suggests that scaled particle theory more precisely estimates the hydrophobic part of solvation entropy than the using a simple surface area term.

Introduction

The estimation of ligand entropy in different environments (vacuum, solution, and protein receptors) is essential for predicting the free energy of ligand transfer between them. While the prediction of entropy of the gas-phase compounds under low and moderate pressures can be achieved using basic statistical thermodynamics expressions for ideal gases,¹ provided that a set of a compound's normal frequencies is given, estimation of that state function in condensed phases is more difficult. The configurational part of entropy, S_c , is given by

$$S_c = -R \int P(\mathbf{r}) \ln P(\mathbf{r}) d\mathbf{r} \quad (1)$$

where R is the gas constant and $P(\mathbf{r})$ is the probability density of the configuration given by coordinates \mathbf{r} , which is often determined from MD simulation of the system where $P(\mathbf{r})$ is derived from the accumulated trajectory² or, assuming that $P(\mathbf{r})$ is a multivariate Gaussian, from the quasiharmonic analysis of the diagonalization of the covariance matrix σ of the coordinate fluctuations:^{3,4}

$$\sigma_{ij} = (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \quad (2)$$

A similar method in which entropy is expressed as a function of coordinate variance, $\langle \Delta r^2 \rangle$, derived from MD simulation was proposed by Schlitter.⁵ All such MD-based methods suffer from the large central processing unit (CPU) times necessary to properly cover phase space, and no matter how long a trajectory is generated, it is always incomplete. Some alternative methods apply corrected versions of the ideal gas-type entropy expressions, particularly to account for finite molecular volume in the translational part of entropy.^{6–8} Although such a correction does indeed eliminate the overestimation of the entropy of a compound in solution upon the usage of ideal gas-type translational expression, in our opinion, it is not well founded because such a reduction is totally accounted for by the entropy effects of solvation phenomena and can be quantitatively described within an adopted solvation model, as, for example, in the recent work of Graziano.⁹ In such a case, the Sackur–Tetrode equation can still be used to estimate the translational part of solute entropy.

These approaches rarely address the issue of conformational entropy. Conformational entropy as a part of configurational ligand entropy was evaluated by Gilson and colleagues^{10,11} and recently applied to protein–ligand binding;¹² however their “mining minima” algorithm requires tens of hours on a commodity computer.

* Corresponding author. E-mail: stan@eyesopen.com.

None of the above approaches are suitable when the evaluation of entropy has to be done for a large number of ligands, for example, during drug design research. There is a need, therefore, for a rapid and reliable method of ligand entropy evaluation.

There are many factors that make the rapid and reliable estimation of ligand entropy in the condensed phase, either in solution or inside a protein receptor, a difficult task. One factor is the conformational diversity of a ligand in solution. Another is the necessity to include external forces acting on ligands in these environments: solvation forces in the case of solution ligands and protein–ligand intermolecular forces for protein-bound ligands. In the latter case, the shape of the protein–ligand potential well limits the motion of the ligand and, therefore, modifies its entropy. Therefore, accurate determination of solvent– and protein–ligand potentials is an important part of ligand entropy calculation.

In this report we present a fast method for estimating ligand entropy based on a Hessian matrix built during optimization with a quasi-Newton optimization. The basic assumption we made is that ligand molecules exist in different conformations in the gas or solution phases and that their fractional numbers in those phases are given by a Boltzmann distribution. This might not be a reasonable assumption for the ligand poses bound in the protein receptor; in this case, we consider a single binding mode given by the crystal structure of the protein–ligand complex.

In the next section, a description of our method is given, followed by its validation against gas- and solution-phase entropies. The last section contains a comparison of experimental and calculated values of $T\Delta S$ for the process of a selected ligand's binding by four protein receptors. We demonstrate that this method is comparable in precision to the more accurate determination of the vibrational part of entropy based on the exact Hessian and is suitable for use in rapid evaluations of ligand entropies.

Methods

Configurational entropy of a ligand in the gas- and solution-phases is calculated from¹

$$S_c = kN \left[1 + \ln \left(\frac{q}{N} \right) + \frac{T}{q} \frac{\partial q}{\partial T} \right] \quad (3)$$

where k is the Boltzmann constant, N is the number of ligand molecules, T is the absolute temperature, and q is the partition function:

$$q = q_t \sum_{i=1}^{n_c} e^{-\varepsilon_i/kT} q_{iv} q_{ir} \quad (4)$$

where q_t is the translational partition function, the summation is over the number of ligand conformers, n_c , q_{iv} , q_{ir} are vibrational and rotational partition functions of conformer i , and ε_i is a sum of the internal energy and solvation free energy of conformation i .

In the case of protein-bound ligands, we assume that three translational and three rotational degrees of freedom of a

ligand are transformed into six degrees of vibrational motion of a trapped ligand, so eq 4 is reduced to

$$q = \begin{cases} q_v & \text{for single binding mode} \\ \sum_{i=1}^{n_p} \exp\left(-\frac{\varepsilon_i}{kT}\right) q_{iv} & \text{for multiple binding modes} \end{cases} \quad (5)$$

where n_p is the number of binding modes and q_{iv} is the vibrational partition function of a bound ligand in mode i .

Translational entropy in solution was calculated from the Sackur–Tetrode equation:

$$S_t = Nk \left(\ln \frac{1}{\rho \Lambda^3} + \frac{5}{2} \right) \quad (6)$$

where $\rho = N/V$ is the number density of the ligand in solution (set at 1 M concentration when molar entropy was evaluated) and Λ is the thermal de Broglie wavelength dependent on the mass of ligand molecule m and temperature:

$$\Lambda = \frac{h}{(2\pi mkT)^{1/2}} \quad (7)$$

No empirical correction to S_t was applied in order to account for the finite volume of solute molecules. The entropic effects of the excluded volume (cavity) not available for the solvent are an important part of solvation entropy and can be explicitly included by adding appropriate solvation terms. We have chosen to use this approach in our calculations (see Solvation Entropy Section). In order to use effectively the above formulation of ligand configurational entropy, we need three fast and reliable computational procedures:

- (i) A method for generating an ensemble of ligand conformations.
- (ii) A method for determining vibrational frequencies of each ligand conformer.
- (iii) Method for estimating solvation effects on solute entropy.

Each method is described briefly below.

Conformation Generation. We have generated conformer ensembles by a method of random coordinates embedding, MMFF94 force field¹³ refinement of fragments, combining of fragments into a molecule and finally torsion driving of rotatable bonds as implemented in Omega (version 2.1).^{14,15} Conformations were generated using default parameters, except an 0.1 Å root-mean-square (RMS) threshold was used to determine and eliminate duplicate conformations. This low limit is intended to assume that the majority of conformers are included in entropy calculations. All conformations generated in this manner were energy minimized with the MMFF94 force field but only structurally unique conformations that differed after minimization by at least 0.05 Å in root-mean-square deviation (RMSD) are included in the entropy calculations. The importance of thorough conformation sampling will be addressed in Results and Discussion Section.

Vibrational Frequencies Determination. Our computational procedure introduces the following approximations: (1) that the vibrational motion of a molecule is represented

as a set of independent, uncoupled oscillators and (2) that each of those oscillators is harmonic. We realize that one might expect that the low-frequency motion of a ligand in the protein binding site might be significantly anharmonic and that the issue of associated error in entropy will need to be addressed in future, but far more important is the quality of the molecular potential that determines the shape of the potential well at minimum and, therefore, the values of the vibrational frequencies. For the purpose of this study, we adopted the MMFF94 potential force field,¹³ which recently has been proved by us to have performed well in the refinement of crystallographic ligand structures when fitted into their electron densities.¹⁶

In order to calculate the vibrational entropy, we derived the vibrational frequencies from the normal-mode analysis of mass-weighted Hessian matrix of second derivatives:

$$\mathbf{H}^m = \mathbf{M}^{-1/2} \mathbf{H} \mathbf{M}^{-1/2} \quad (8)$$

where \mathbf{H} and \mathbf{M} are the Hessian and atomic mass matrices, respectively. Eigenvalues λ_i of \mathbf{H}^m determine harmonic wavenumbers $\tilde{\nu}_i$ following eq 9:

$$\tilde{\nu}_i = \frac{1}{2\pi c} \sqrt{\lambda_i} \quad (9)$$

where c is the speed of light.

When the molecular potential is fully analytic, the Hessian matrix can be calculated for the optimized ligand structure. However, in order to make calculations more efficient and extend them for cases where some potential terms are given in the form of a high-resolution lookup table or three-dimensional (3D) grid, this work uses a Hessian matrix built by quasi-Newton type optimizers for the purpose of predicting the next step:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \mathbf{H}_{i+1}^{-1} \nabla f(\mathbf{x}_i) \quad (10)$$

where \mathbf{x}_i is the coordinate vector at step i , \mathbf{H}_{i+1}^{-1} is an inverse of the Hessian matrix at step $i + 1$ constructed according to the adopted scheme (for example, the widely used Broyden–Fletcher–Goldfarb–Shanno, BFGS, scheme),¹⁷ and $f(\mathbf{x})$ is a function for which minimum is sought. At every iteration the approximate \mathbf{H}^{-1} is closer to the exact Hessian, so when the optimization procedure is converged, a good quality Hessian can be obtained. Before the normal frequencies are calculated, we make sure that the Hessian is stable and does indeed determine a real minimum. If one or more eigenvalues are negative, the geometry of the ligand is randomly perturbed and reoptimized using the final matrix \mathbf{H}^{-1} from the previous optimization as an initial guess of the invert Hessian. We have found that such a correction is needed very rarely when the molecular potential is fully analytic and, more frequently but still only occasionally, when some potential terms are evaluated on a grid.

The procedure outlined above is expected to result in frequencies and moments of inertia reasonably close to experiment only for the gas phase or for small ligands in solution with no rotatable bonds. The structure of highly flexible polar ligands, however, may be largely changed in solution. That implies a need to include solvent forces in

order to properly estimate vibrational and rotational entropy in solution. Here, a simple analytical solvation model recently developed by Grant and colleagues, known as the Sheffield solvation model,¹⁸ is adopted. Briefly, the solvation energy of a ligand in solution is expressed as

$$E_s = -\frac{f_\epsilon}{8\pi\epsilon_0} \sum_{ij} \frac{Q_i Q_j}{\sqrt{a r_i r_j + b R_{ij}^2}} \quad (11)$$

where ϵ_0 is permittivity of vacuum, Q_i and Q_j are partial charges on atoms i and j , r_i and r_j are atomic radii, R_{ij} is the distance between atoms i and j , and the factor f_ϵ is defined by dielectric constants of the solvent and the ligand (solute):

$$f_\epsilon = \left(\frac{1}{\epsilon_{\text{solv}}} - \frac{1}{\epsilon_{\text{solv}}} \right) \quad (12)$$

Dimensionless parameters a and b have been chosen in such a way that the solvation energy given by eq 11 agrees with a physically rigorous model based on the Poisson equation.

Solvation Entropy. Solvation of a ligand in solution is associated with the formation of cavities around solute molecules and the reorganization of water molecules around them due to electrostatic and nonelectrostatic solute–solvent interactions. Each of those phenomena are accompanied by entropy changes. Comparison of estimated ligand entropy in solution with that of the corresponding experimental data, therefore, requires inclusion of solvation entropy ΔS_s (along with solute configurational entropy S_c) in the expression for the total ligand solution entropy:

$$S_{\text{solution}} = S_c + \Delta S_s \quad (13)$$

ΔS_s consists of electrostatic and hydrophobic parts:

$$\Delta S_s = \Delta S_{s,\text{elec}} + \Delta S_{s,\text{hyd}} \quad (14)$$

The way we estimate both parts of solvation entropy is given below.

Electrostatic Solvation Entropy. Bulk effects of entropy change upon solvation due to the electrostatic properties of the solution result from the temperature dependence of the solvent dielectric constant. This portion of solvation entropy is estimated from

$$\Delta S_{s,\text{elec,bulk}} = -\left(\frac{\partial \Delta G_s}{\partial \epsilon_{\text{solv}}} \right) \left(\frac{\partial \epsilon_{\text{solv}}}{\partial T} \right) \quad (15)$$

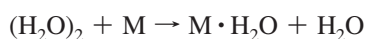
where ϵ_{solv} is the solvent dielectric constant. The first term in eq 15 can be calculated from the simple analytical solvation model, while the second term is calculated from the experimental temperature dependence of the water dielectric constant.¹⁹

Entropic effects that reflect the change of the water structure upon solvation, in particular formation of hydrogen bonds between the solute and the solvent, are difficult to estimate and are beyond the methods that treat water as dielectric continuum. Water is a highly structured liquid in which every H_2O molecule with two proton donor and two proton acceptor sites forms four hydrogen bonds (H-bonds)

with its neighbors in a tetrahedral arrangement. Entropy change for the formation of such a single H-bond in water is estimated to be -37.2 J/(mol K).²⁰ Solute molecules containing H-bond donors or acceptors engage water molecules from the first solvation shell in solute–solvent H-bond formation, thus perturbing the original highly organized liquid structure. Also, the vibrational entropy of the H-bonded solute·(H₂O)_n complexes is usually higher than that of the corresponding water clusters due to the larger number of low-frequency modes available in the former.

In order to account for the formation of ligand–solvent clusters due to strong hydrophilic interactions, we propose using of an empiric constant term of 28 J/(mol K) that increases the solution entropy of a compound containing proton donors or acceptors. The rationale of such a choice is based on the following two observations:

- (i) Calculations of the vibrational entropy change for a number of gas-phase reactions:



for a number of molecules M that contain proton donors or acceptors with the use of the MMFF94 potential yield a vibrational entropy change in the range 15 – 50 J/(mol K). One might expect, therefore, a small favorable entropic effect enhancing solubility of such compounds.

- (ii) Our calculations of solution entropy for polar molecules or molecules with π -electron donors produce consistently underestimated values by 10 – 45 J/(mol K), with the remarkable exception of alcohols, for which a satisfactory agreement between calculated and experimental solution entropies is observed. One explanation for this observation is that in all cases but alcohols we are missing the entropy change of the H-bond rearrangements upon solvation in the first solvation shell. It is possible that the hydroxyl group in alcohols, with a very similar charge distribution to the hydroxyl group in water, does not significantly perturb the water structure even in the first solvation shell, while the increase of the vibrational entropy upon ROH·H₂O formation is small.

The proposed entropy correction accounting for solute–solvent clusters formation due to specific hydrophilic interactions is certainly a crude approximation, however, its accurate calculation for drug-like ligands is too difficult at this stage of our method development.

Hydrophobic Solvation Entropy. Calculation of the hydrophobic portion of solvation entropy is usually performed by one of a few approximate methods. We estimate this solvation entropy term in two ways. One is based on the common assumption that the free energy of hydrophobic solvation is proportional to the solute molecular surface area

$$\Delta G_{\text{s,hyd}} = \gamma A \quad (16)$$

where γ is the microscopic surface tension coefficient. The value of γ is in the range of 0 – 10 cal/(mol Å)^{21,22} when $\Delta G_{\text{s,hyd}}$ represents the overall hydrophobic solvation effect, which includes both cavitation and van der Waals solute–

solvent interactions. We may, therefore, assume that γ is made of two components: $\gamma = \gamma_{\text{cav}} + \gamma_{\text{vdw}}$. Assuming that $\Delta G_{\text{s,hyd}}$ depends linearly on temperature and that its enthalpic contribution does not depend on temperature, the corresponding entropy change is

$$\Delta S_{\text{s,hyd}} = -\frac{\Delta G_{\text{s,hyd}}}{T} \quad (17)$$

Under the above assumptions, $\Delta G_{\text{s,hyd}}$ in eq 17 represents the temperature-dependent cavity formation term for which the value of $\gamma = \gamma_{\text{cav}}$ is about 30 cal/(mol Å), as determined for alkanes.^{23,24} Using a set of 294 molecules containing a variety of functional groups, we have found that $\gamma = 30$ cal/(mol Å) indeed returns the best agreement on average with solution entropies.

The main criticism of using eq 16 is that for small molecules the hydrophobic portion of entropy is primarily related to the creation of a cavity in the solvent and scales with its volume²⁵ rather than the molecule's surface area A . In addition, coefficient γ has no precisely established value for all compounds. One has to conclude, therefore, that the widely used expression (eq 16), although useful, is not physically sound.

For the purpose of comparison, we apply, therefore, an alternative approach in which the hydrophobic free energy change $\Delta G_{\text{s,hyd}}$ is evaluated as a sum of a cavity formation component, van der Waals solute–solvent interaction and inductive (permanent dipole–induced dipole) interaction:

$$\Delta G_{\text{s,hyd}} = \Delta G_{\text{cav}} + \Delta G_{\text{vdw}} + \Delta G_{\text{ind}} \quad (18)$$

The first component in eq 18 is evaluated from the scaled particle theory (SPT)^{9,26,27} according to the expression containing up to the cubic term:

$$\Delta G_{\text{cav}} = RT[K_0 + K_1(\sigma_c/\sigma_s) + K_2(\sigma_c/\sigma_s)^2 + K_3(\sigma_c/\sigma_s)^3] \quad (19)$$

where $K_0 = -\ln(1 - \xi)$, $K_1 = 3\xi/(1 - \xi) = u$, $K_2 = u(u + 2)/2$, $K_3 = \xi P v_s/RT$, σ_c and σ_s are cavity and solvent diameters, respectively (where solvent and solute are assumed spherical), ξ is the packing density of the solvent, and v_s is the solvent molar volume. The second term in eq 18 is calculated according to Pierotti's relationship:²⁶

$$\Delta G_{\text{vdw}} = -(64/3)\xi\varepsilon_{\text{ls}}(\sigma_{\text{ls}}/\sigma_c)^3 \quad (20)$$

where ε_{ls} is Lennard-Jones (LJ) potential depth parameter for ligand–solvent interaction calculated as $(\varepsilon_l\varepsilon_s)^{1/2}$ where ε_l and ε_s are the corresponding parameters for ligand and solvent, respectively. The last term in eq 18, in most cases very small (of the order of 0.1 J/(mol K)) in comparison to ΔG_{cav} and ΔG_{vdw} , was evaluated according to²⁶

$$\Delta G_{\text{ind}} = -8\xi(\sigma_c\sigma_{\text{ls}})^{-3}(\mu_l^2\alpha_s + \mu_s^2\alpha_l) \quad (21)$$

where μ_s , α_s , μ_l , and α_l are solvent and ligand dipole moment and polarizability, respectively. The value of σ_{ls} is taken as the average $(\sigma_s + \sigma_c)/2$. $\Delta S_{\text{s,hyd}}$ is determined from

$$\Delta S_{s,\text{hyd}} = \Delta S_{\text{cav}} - \left(\frac{\partial \Delta G_{\text{vdw}}}{\partial T} + \frac{\partial \Delta G_{\text{ind}}}{\partial T} \right) \quad (22)$$

where the entropy change for cavity formation ΔS_{cav} is calculated from ΔG_{cav} (eq 19) and the corresponding enthalpy change

$$\Delta H_{\text{cav}} = [\xi \alpha RT^2 / (1 - \xi)^3] [(1 - \xi)^2 + 3(1 - \xi)(\sigma_c / \sigma_s) + 3(1 + 2\xi)(\sigma_c / \sigma_s)^2] + \xi P v_s (\sigma_c / \sigma_s)^3 \quad (23)$$

as

$$\Delta S_c = (\Delta H_{\text{cav}} - \Delta G_{\text{cav}}) / T \quad (24)$$

Temperature derivatives of ΔG_{vdw} and ΔG_{ind} are determined from the known experimental temperature dependence of σ_s , ξ , and v_s for water. Quantity α is the thermal expansion coefficient of the solvent.

The accuracy of all the thermodynamic quantities derived from the SPT theory (eqs 19–24) depends critically on solvent parameters. Values adopted in this work for water are: molecular volume $v_s = 18.0685 \text{ cm}^3 \text{ mol}^{-1}$, thermal expansion coefficient $\alpha = 2.572 \times 10^{-4} \text{ K}^{-1}$, effective hard sphere diameter $\sigma_s = 2.8 \text{ \AA}$, LJ potential depth $\epsilon_s/k = 100 \text{ K}$, dipole moment $\mu_s = 1.8 \text{ D}$, and polarizability $\alpha_s = 1.4573 \text{ \AA}^3$. The value of the hard sphere diameter for water of 2.8 \AA results from the location of the first peak in the oxygen–oxygen radial distribution function for water,²⁸ while the LJ parameter of 100 K (ratio of Lennard-Jones parameter ϵ_s to Planck's constant, to be correct) was adopted after Graziano⁹ as the value that works well for alkanes and alcohols.

In the case of solute parameters, we applied the simple procedures described below to estimate all necessary parameters (σ_c , ϵ/k , μ_1 , and α_1). For molecules with multiple conformations, we used the geometry of the most stable conformation in solution optimized with the MMFF94 and Sheffield forces.

Hard Sphere Diameters. For each solute, the hard sphere parameter was calculated from its molecular Gaussian volume V_G^{29} as $2(0.75V_G/\pi)^{1/3}$. We found a fair agreement between hard sphere diameters calculated in this way with the values obtained from solubility data for a number of molecules listed in Table 1.

LJ Potential Depth. Experimental LJ potential parameters derived from either second virial coefficients or viscosity measurements are available only for a small number of molecules. Critically evaluated data based on the second type of measurements are available for 75 simple molecules in the 1977 paper of Mourits and Rummens.³¹ Two decades later, Cuadros et al.³² developed a simulation-based procedure for evaluating LJ parameters for any small compound, which was recently used by Cachadina and Mulero for calculating the vaporization enthalpies for over 1500 substances.³³ Our procedure of assigning the LJ potential depth parameter of a compound uses both above-mentioned sources of data in a hierarchical fashion: if the compound happens to belong to the Mourits and Rummens set, the corresponding experimental value is used; otherwise the data obtained by Cachadina and Mulero are adopted. If a compound is not

found in any of the above two sets, an average value of 370 K for σ_c/k is used. This represents the average of the Cachadina and Mulero set excluding mono-, di-, and triatomic species.

Dipole Moments. The dipole moment for the most stable solution conformation calculated from its AM1BCC partial charges was used. Because such a value usually overestimates the experimentally measured dipole moment, we further scale it by a factor of 0.82 , which is the ratio of experimental-to-calculated dipole moment for water molecule. Other molecules are known to have similarly overpolarized charges with the AM1BCC charging method.

Molecular Polarizabilities. The molecular polarizability of a compound was estimated according to the recently published method of Wang et al.³⁴ (model 2E in Table 4 in that publication).

Results and Discussion

When conformation ensembles are used for the estimation of entropy, an important issue is the completeness of those ensembles. The lack of thorough conformation sampling can be illustrated with the gas-phase *n*-nonane: using an RMS threshold of 0.8 \AA for duplicate removal results in 25 conformations, which, after minimization, appear to be unique. The total calculated entropy of *n*-nonane at 298 K for this set of conformations is $452.9 \text{ J}/(\text{mol K})$. The use of a 0.1 \AA RMS threshold produces 130 conformations which after minimization are reduced to 128 structurally unique conformers and results in $472.7 \text{ J}/(\text{mol K})$ total entropy. The error of about $20 \text{ J}/(\text{mol K})$ (or 4%) can easily be eliminated by a fine-grain sampling of the conformational space. Further lowering of the duplicate removal threshold to an RMS of 0.005 \AA has no effect: 239 generated conformations contained only 128 structurally unique conformers after optimizations.

Gas-Phase Molecules. Gas-phase entropies calculated with the MMFF94 force field and quasi-Newton Hessian frequencies for the vibrational entropy component for most small molecules with no torsions or containing only an isolated methyl group bonded to aromatic rings (like toluene) are in very good agreement with experimental values. This is shown in Figure 1a for a number of molecules containing carbon, oxygen, sulfur, nitrogen, and halogen atoms. In contrast, molecules with single torsional bonds show up to 8% error in calculated entropies with respect to experimental values. This behavior is shown in Figure 1b for *n*-alkanes from ethane to decane and is not surprising given the underlying approximations described in the previous section. In particular, we might expect that the torsional, low-frequency vibrations are significantly anharmonic, so the error in entropy estimation for larger, more flexible molecules will be larger than for rigid ones. A good illustration of poorer entropy estimation for flexible and satisfactory predictions for rigid molecules are calculated values for benzene and *n*-hexane of 270.2 and $367.3 \text{ J}/(\text{mol K})$, respectively, and the corresponding experimental values of 269.2 and $388.4 \text{ J}/(\text{mol K})$. Similarly, the RMS deviation between calculated and experimental values in the series of

Table 1. Comparison of Hard Sphere Diameters Derived from Gaussian Molecular Volumes (σ_G)^a

compound	σ_G	σ_{solu}	compound	σ_G	σ_{solu}
carbon dioxide	4.10	3.94	benzene	5.29	5.26
methane	3.60	3.70	toluene	5.61	5.64
ethane	4.20	4.38	<i>m</i> -xylene	5.89	5.97
ethylene	4.07	4.07	fluorobenzene	5.34	5.30
<i>n</i> -hexane	5.73	5.92	chlorobenzene	5.71	5.61
<i>n</i> -heptane	6.01	6.25	hexafluorobenzene	5.59	5.65
<i>n</i> -octane	6.26	6.54	nitrobenzene	5.85	5.74
<i>n</i> -nonane	6.50	6.83	methanol	4.01	3.69
<i>n</i> -decane	6.71	7.08	ethanol	4.53	4.34
3-methylheptane	6.26	6.52	cyclohexanol	5.76	5.75
2,3-dimethylhexane	6.26	6.50	acetone	4.86	4.76
cyclohexane	5.57	5.63	<i>N</i> -methylacetamide	5.15	4.96
methylcyclohexane	5.86	5.99	dimethylsulfoxide	5.10	4.91

^a Corresponding values obtained from solubility data by Wilhelm and Battino³⁰ (σ_{solu}). All values in Å.

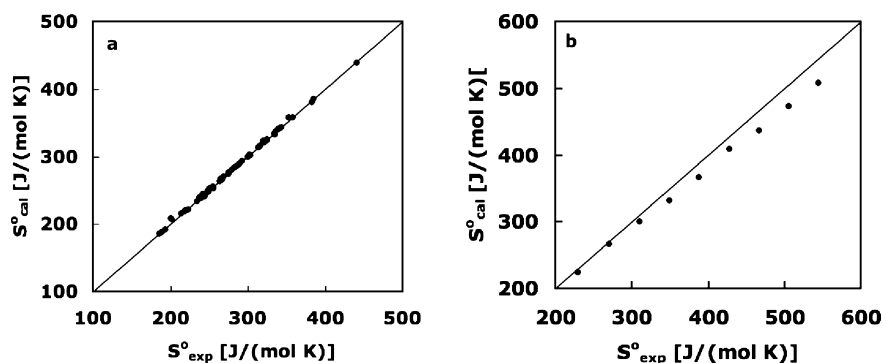


Figure 1. Calculated versus experimental entropies at 298K for the below-listed gas-phase molecules. Plot a: methane,^a ethylene,^a acetylene,^a cyclopropane,^b allene,^b benzene,^b toluene,^b naphthalene,^b water,^a carbon dioxide,^a formaldehyde,^a formic acid,^a oxirane,^a methanol,^b phenol,^b thiacyclopropane,^b thiophene,^b methanethiol,^b ammonia,^a hydrogen cyanide,^a aziridine,^b pyridine,^b nitric acid,^a nitrous acid,^a acetonitrile,^b benzonitrile,^b nitromethane,^b methyl nitrate,^b 3-picoline,^b fluoromethane,^a tetrafluoroethylene,^b 1,1-difluoroethylene,^b trifluoroethylene,^b fluorobenzene,^b fluorotoluene,^b *p*-fluorotoluene,^b chloromethane,^a chloroethylene,^b tetrachloroethylene,^a 1,1-dichloroethylene,^b *t*-1,2-dichloroethylene,^b trichloroethylene,^b chlorobenzene,^b 1,2-dichlorobenzene,^b 1,3-dichlorobenzene,^b 1,4-dichlorobenzene,^b hexachlorobenzene,^b bromomethane,^a bromoethylene,^b bromomethane,^b chlorotrifluoroethylene,^b iodomethane,^b iodobenzene.^b Plot b: *n*-alkanes from ethane to *n*-decane. ^aExperimental values are taken from ref 35, and ^bexperimental values are taken from ref 36. Solid diagonal lines show ideal behavior of calculated values.

thiacycloalkanes from thiacyclopropane to thiacycloheptane is 5.7 J/(mol K), while the corresponding value for 1-thiols from ethanethiol to 1-hexanethiol is 22.4 J/(mol K).

Anharmonicity error in the calculation of normal frequencies has been recognized for a long time in quantum chemical calculations; a number of remedies, including scaling frequencies by factors within the (0.8, 1.0) range,³⁷ selective or overall scaling of force constants,³⁸ and estimation of anharmonicity constants by numerical calculation of third and fourth potential derivatives along the normal modes,^{39,40} have been proposed and successfully used to predict the gas-phase IR and Raman frequencies of a variety of small molecules. For the purpose of our method, we adopted a simple frequency scaling. Based on a set of 255 molecules for which standard gas-phase entropies are available in the compilation of Domalski and Hearing,³⁶ we obtained an optimum scaling factor of 0.85. Figure 2, compares calculated gas-phase standard entropies with scaled and unscaled frequencies with the experimental (or recommended) values for the above-mentioned set of 255 molecules. This figure shows that scaling of frequencies results in much better agreement with the published experimental or recommended entropies. Unfortunately, for some molecules the scaling

Table 2. Comparison of Molecular Diameters Obtained in Two Ways^{a,b}

compound	experimental	unscaled	scaled
cyclohexanone	322.2	341.1	354.5
thiacyclopentane	361.9	371.3	389.9
hexachlorobenzene	441.2	439.1	463.8

^a σ_G - hard sphere diameters derived from Gaussian molecular volume, and σ_{solu} - diameters obtained from the solubility data by Wilhelm and Battino.³⁰ ^b All values in Å.

procedure results in significant overestimation of entropy that is particularly visible in the range of 300–500 J/(mol K). For example, visibly deviated points from the ideal line correspond to cyclohexanone, thiacycloheptane, and hexachlorobenzene, for which the experimental and calculated standard entropies in J/(mol K) with unscaled and scaled normal frequencies are see in Table 2.

It is seen that for these compounds, frequency scaling increases the error by largely overestimating the calculated entropy.

Another source of error is the approximate character of the quasi-Newton Hessian matrix. As mentioned in the Vibrational Frequencies Determination Section, a good

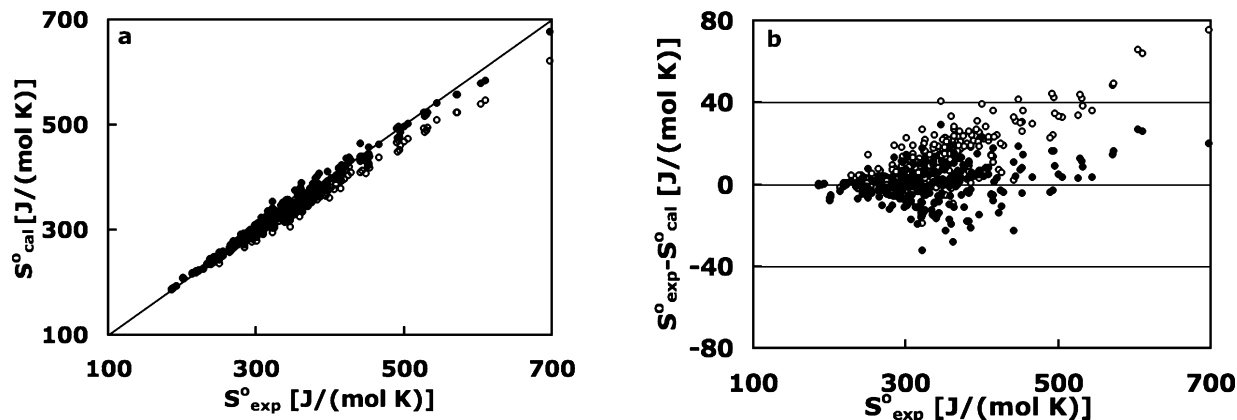


Figure 2. Gas-phase standard entropies (plot a) and their displacements from experimental values (plot b) calculated with the use of unscaled (○) and scaled (●) normal frequencies for a set of 255 molecules containing carbon, oxygen, sulfur, nitrogen, and halogen atoms. A list of molecules other than those mentioned in the caption of Figure 1 is given in the Appendix Section. The RMS displacement from published entropies is 18.2 and 9.4 J/(mol K) for unscaled and scaled normal frequencies, respectively.

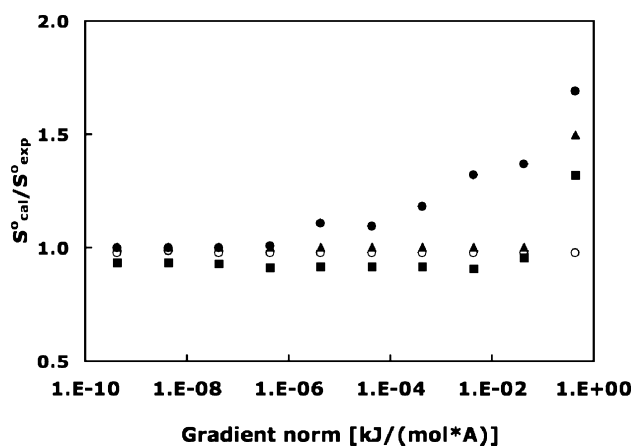


Figure 3. Ratio of calculated-to-experimental gas-phase entropy for ethane (○), *n*-decane (□), benzene (▲), and cyclohexane (●) as a function of gradient norm convergence criteria in quasi-Newton built Hessian.

quality approximation of the quasi-Newton Hessian matrix is obtained only when the optimization process is converged. Figure 3 shows the effect of convergence criteria on the calculated gas-phase entropy of four hydrocarbon molecules. It is seen that, although for a molecule with no torsions, like benzene, gradient norm reduction to 10^{-3} kJ/(mol Å²) is sufficient to obtain a stable converged value of entropy, the data for cyclohexane suggest that a convergence criteria of at least 10^{-7} kJ/(mol Å²) is necessary to minimize the error related to the approximated Hessian obtained in the quasi-Newton optimization. Therefore, in all entropy calculations we use a convergence on gradient norm below that value, typically 10^{-10} kJ/(mol Å²).

In order to estimate the effect of the approximate character of the quasi-Newton converged Hessian, we repeated the calculations of entropies for the same set of 255 compounds from Figure 2 but with the exact, analytical Hessian matrix calculated for the optimized compounds. The RMS displacement of entropies calculated in such a way from published values is 15.0 J/(mol K), compared to 20.8 J/(mol K) obtained with the quasi-Newton frequencies. The improvement due to the diagonalization of the exact Hessian is,

therefore, real (as might be expected) but on average is not dramatic. The quality of the force field is probably of more significance, but in this paper, we make no attempt to evaluate different available molecular potentials for entropy estimation.

Molecules in Solution. The goal of this study is to elucidate a very rapid means of calculating ligand entropies in gas and condensed phases. Thus, we sought an approximate solvation model that could rapidly evaluate solvent forces that modify equilibrium geometries and normal frequencies of ligand conformations in solution. To determine the accuracy of ligand entropies calculated with this approximate solvent model, we compared the values calculated with a more rigorous but significantly slower Poisson model. Calculated configurational entropies for 20 drug molecules in solution with the use of a simplified analytical solvation Sheffield model¹⁸ in comparison to a Poisson model are shown in Figure 4a. It is seen that for the majority of molecules, the Sheffield values are slightly larger. The average signed error of Sheffield entropies is 13.9 J/(mol K). Observed differences might also, however, be impacted by the larger numerical errors in the calculation of the quasi-Newton Hessian in the case of the Poisson potential evaluated on the grid with respect to the fully analytical Sheffield potential. Figure 4b compares the CPU times of both types of entropy calculations and shows the obvious speed advantage in the case of Sheffield model. Large CPU times visible on the in Figure 4b for the Poisson-type entropy calculations reflect numerous reoptimizations (as outlined in the Methods Section) needed for some conformations to meet the formal requirements for a Hessian matrix. Differences in configurational entropy calculated with the two solvent models are small (3–6%), yet on average, the Sheffield model is over two orders of magnitude faster to calculate. We conclude that using the Sheffield solvation model for rapid entropy estimation of solution molecules is a reasonable and beneficial approach.

Total solution entropies calculated according to eq 13 can be compared with the corresponding experimental entropy

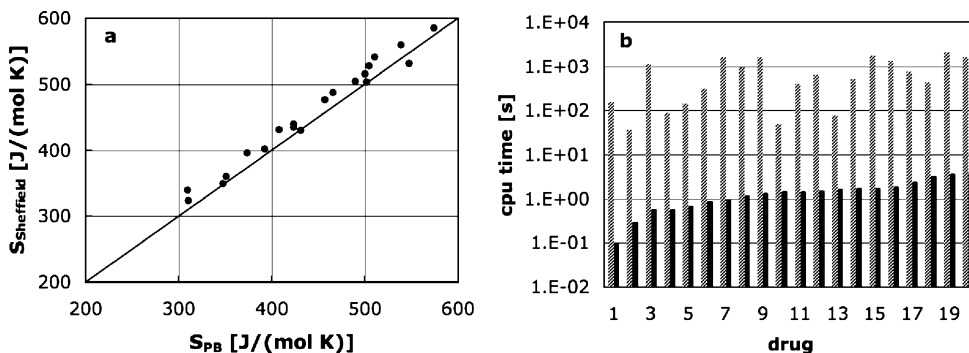


Figure 4. (a) Calculated configurational entropies for 20 drug molecules using Sheffield solvation model ($S_{\text{Sheffield}}$) vs Poisson model (S_{PB}). (b) CPU times used for entropy calculations for each drug using Sheffield model (solid bars) and Poisson model (shaded bars). Numbers on horizontal axis correspond to the following drugs: varenicline (1), acetaminophen (2), aspirin (3), ephedrine (4), telbiduvine (5), lamivudine (6), ofloxacin (7), decitabine (8), meloxicam (9), hydrocortizone (10), sertraline (11), ciproflaxin (12), desonide (13), ibuprofen (14), zoledronic acid (15), neralabine (16), ritalin (17), venlaxafine (18), warfarin (19), and tramadol (20).

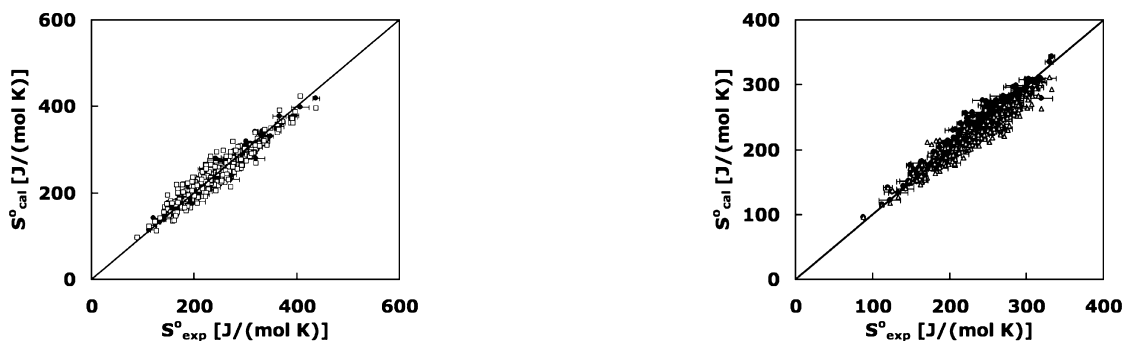


Figure 5. Calculated vs experimental standard molar entropies for 294 molecules in solution representing the following group of compounds: normal and branched alkanes, cycloalkanes, alkenes, alkynes, alkylbenzenes, alcohols, ketones, ethers, esters, thiols, sulfides, nitriles, and amines. Experimental data are taken from the ORCHYD database⁴¹ except those for amines and three purines (adenine, xanthine, and hypoxanthine). Experimental data for amines are taken from the 1981 paper of Cabani et al.⁴² Data for purines are taken from papers of Boyer et al.⁴³ and Tewari et al.⁴⁴ Full circles (●) show calculated standard entropies with the use of the SPT theory, while squares (□) show similar data calculated with the SA eqs 16 and 17 with tension coefficient γ set at 30 cal/(mol/Å). The values of R^2 and slopes (R^2 , slope) are (0.945, 0.923) and (0.878, 0.865), respectively. Corresponding lower and upper 95% confidence intervals for R^2 , obtained from Fisher transformation are (0.931, 0.956) for SPT and (0.849, 0.902) for SA solvation models, respectively. Straight line shows ideal agreement with experiment.

data. Figure 5 shows such a comparison for 294 molecules containing a variety of functional groups.

It is seen in Figure 5 that the calculated values follow the experimental values; however, the use of the SPT theory returns a better agreement with experimental data, as is visible from slopes, R^2 's, and confidence intervals for R^2 . The root-mean-square error (RMSE) of the calculated standard solution entropies using the SPT model from the corresponding experimental data is 13.6 J/(mol K), while the largest deviation visible on the plot is the underestimation of the molar solution entropy by 39.8 J/(mol K) for 1,8-nonadiyne. The corresponding values for the SA model are 20.0 and 56.0 J/(mol K) overestimation for trans-1,4-dimethylcyclohexane. The differences in prediction accuracy

Figure 6. Calculated vs experimental standard molar entropies for 244 molecules, a subset of molecules for which data are presented in Figure 5. Calculations were done with the SPT/Sheffield (●) and SPT/Poissson (△). The corresponding values of R^2 and slopes are (0.911, 0.906) and (0.899, 0.808), respectively. The 95% confidence intervals for R^2 are (0.887, 0.930) and (0.872, 0.921), respectively.

between the two models are not dramatic but statistically significant: larger R^2 for the SPT method (0.95 vs 0.88) together with the lack of overlap of 95% confidence intervals for R^2 , shown in the caption for Figure 5, clearly illustrates that the SPT model outperforms the SA method in prediction of solution entropies. We should also mention that our SPT method has no adjustable parameters, whereas the SA method was optimized albeit to a value that could have been estimated from the physical properties of alkane–water transfer.

Figure 6 shows the comparison between total solution entropies calculated with Sheffield/SPT and physically more rigorous PB/SPT model. There is little difference in the use of the Sheffield or the full PB model; while the slopes are slightly different, the correlation coefficients are identical within statistical error of 95% confidence. This is a remarkable result, indicating that a simple, fully analytical, and fast solvation model generates the same quality entropy prediction for molecules in solution that the physically correct but much more expensive to use Poisson solvation model. As discussed above, slightly better prediction of Sheffield/SPT method is not surprising and can be attributed to the higher accuracy of analytical second-order derivatives.

We hope that using a molecular potential of higher quality than MMFF and refinement of the model describing solvent–solute interactions contribution to the solvation

terms (both vdw and hydrogen bonding) (eq 14) could lead to a better agreement with experimental data. It was shown that, with water as a solvent, SPT provides reliable results for simple small molecules,^{45,46} however, we are not aware of its prior application to complex drug-like compounds. An extended version of SPT for solutes of arbitrary shapes has been developed,⁴⁷ but it is not clear if it generates significantly better results in aqueous solutions than its original hard-sphere version, so at this point, we do not have plans to implement it just for the purpose of entropy calculations.

Protein-Bound Molecules. The only way to test our method of estimated ligand entropy inside a protein receptor is to compare calculated and experimental entropy change ($T\Delta S$) upon ligand binding. Only limited amounts of such experimental data from microcalorimetry experiments exist. In addition, cases where there is a significant contribution to the overall entropy change from the protein itself, caused by large change in protein structure upon binding, cannot be used for the purpose of testing because we are not attempting here to estimate protein entropy change. In fact, our basic approximation in evaluating a protein-bound ligand is the complete rigidity of a protein. Such a crude approximation, therefore, severely restricts the scope of protein–ligand complexes which might be handled by our method.

We have chosen four protein–ligands systems shown below. In selecting these systems, our main criteria were lack of favorable entropy change upon binding ($T\Delta S > 0$) (protein contribution to $T\Delta S$ cannot be ignored otherwise), moderate size of ligands, quality of experimental data, and of course, existence of the protein–ligand X-ray structures in the Protein Data Bank (PDB):⁴⁸

- (i) Major urinary protein (MUP-I)–*n*-alcohols (1znd, 1zne, 1zng, 1znh, and 1znk).
- (ii) Renin–diaminopyrimidines (2iko, 2iku, and 2il2).
- (iii) Aldose reductase–sorbitin/fidarestat (2pdk and 1pwm).
- (iv) Quinone reductase–resveratrol/melatonin (1sgo and 2qx4).

X-ray protein structures used for calculations are shown above. Protein preparation included hydrogenation to standard residues ionization states and optimization of hydrogen atom positions with the MMFF94 force field in vacuum. Entropy change upon ligand binding has to include partial solvation of that part of the ligand in the protein–ligand complex which is exposed to the solvent, $f\Delta S_s$, and partial desolvation of the protein binding site, ΔS_{des} , upon ligand binding.

$$\Delta S_{bind} = S_{protein} - S_{solution} + f\Delta S_s + \Delta S_{des} \quad (25)$$

ΔS_s is given by eq 14, and the fraction f of a bound ligand surface exposed to the solvent is evaluated from the molecular surfaces area of a ligand A_L , protein A_P , and protein–ligand complex A_{PL} :

$$f = 0.5(A_L - A_P + A_{PL})/A_L \quad (26)$$

Entropy of partial protein desolvation, ΔS_{des} , is assumed to be caused by the change of the hydrophobic part of protein desolvation and is calculated from the surface area expression (eq 17) because for macromolecules it scales with the molecular surface.²⁵ The microscopic surface tension coefficient γ is set at 5 cal/(mol Å).

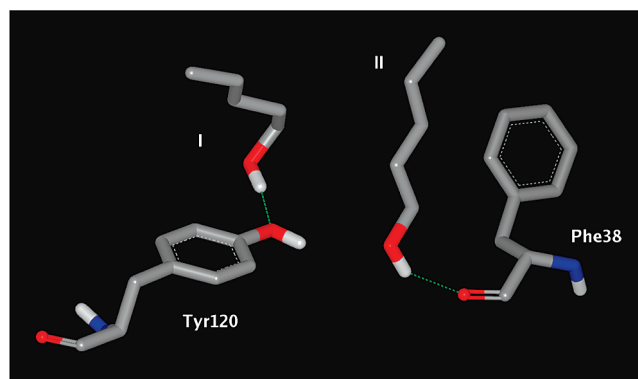
MUP-I–*n*-Alcohols. Structures for the complexes of pheromone-binding protein MUP-I with five primary alco-

Table 3. Calculated and Experimental Values of $-T\Delta S$ for the Binding Equilibria MUP-I + *n*-Alcohol \rightleftharpoons MUP-I·*n*-Alcohol^a

alcohol	$-T\Delta S_{cal}$	$-T\Delta S_{exp}$
pentanol	18.4, 22.6 ^b	17.9 ± 3.2
hexanol	32.3, 21.0 ^b	19.3 ± 0.6
heptanol	30.6, 19.9 ^b	20.9 ± 0.4
octanol	25.7	22.4 ± 0.6
nonanol	26.1	24.8 ± 0.5

^a $-T\Delta S$ in kJ/mol. ^b Value for the alternative position. ^c Value for double binding modes.

hols: pentanol, hexanol, heptanol, octanol, and nonanol were taken from the PDB entries 1znd, 1zne, 1zng, 1znh, and 1znh, respectively. The protein structures in these models are essentially unchanged in each complex, but ligands can be bound in two structurally different positions depicted on Structure I. In the case of the pentanol complex, Malham et al. have observed both positions simultaneously, however, the second one (mode II) with much weaker density.⁴⁹ In the first run of entropy calculations, we assumed a single binding mode with starting geometries given in the above crystal structures. Table 3 contains the results along with the experimental results of isothermal titration microcalorimetry (ITC) performed by Malham et al.⁴⁹ It is seen that the calculated entropy penalty agrees very well with the experimental values for pentanol and fairly well with the corresponding values for octanol and nonanol. For hexanol and heptanol, our prediction is too negative by 13.0 and 9.7 kJ/mol, respectively. In order to explain the difference, we repeated the calculations for those two alcohols and for pentanol, assuming that the initial position for pentanol is given by the alternative position observed experimentally by Malham et al.,⁴⁹ while the starting poses for hexanol and heptanol were obtained by truncating the terminal carbons from the octanol 1znh structure (mode II in structure I). Recalculated $-T\Delta S$ values for pentanol, hexanol, and heptanol are marked in Table 3 with a footnote (b, in this case). It is seen that those latter values for both hexanol and heptanol are in very good agreement with the experimental data of Malham et al.⁴⁹ In the case of pentanol, the alternative pose leads to larger entropy penalty and larger deviation from the experimental value (-22.6 vs -18.4 kJ/mol).



Structure I

Our calculations suggest that as alcohol molecules grow, due to steric hindrance only, mode II seems to be available,

Table 4. Calculated and Experimental Values of $-T\Delta S$ for Several Binding Equilibria Protein + Ligand \rightleftharpoons Protein·Ligand^a

protein	compound	$-T\Delta S_{\text{cal}}$	$-T\Delta S_{\text{exp}}$
renin	I	12.6	8.4 ± 4.2
	II	7.5	-1.0 ± 4.2
	III	3.3	1.8 ± 4.2
aldose reductase	IV	11.7	13.9 ± 1.6
	V	13.1	28.8 ± 0.8
	IV ^b	19.7	
	V ^b	19.9	
quinone reductase 2	VI	23.1	21.0 ± 2.6
	VII	10.0	10.0 ± 0.5

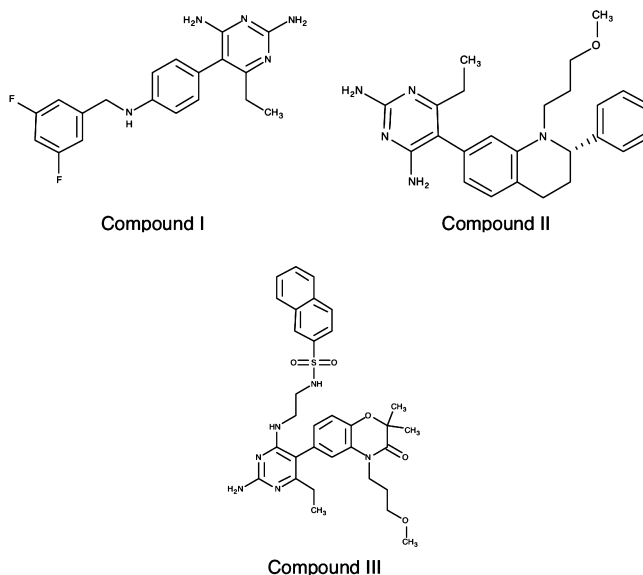
^a Reported experimental error for renin binding are from a personal communication of R.W. Sarver. $-T\Delta S$ in kJ/mol. ^b Anion bound by a protein with protonated His110.

while for smaller pentanol, both are likely to occur with the majority binding in mode I. X-ray data for pentanol, octanol, and nonanol complexes are in full agreement with the above suggestion. The source of disagreement for hexanol and heptanol is not clear. One possibility is that our model does not include explicit water molecules in the binding site. Such water molecules were observed in X-ray data by Malham et al.,⁴⁹ and because they form H-bonds with hydroxyl groups of both bound alcohols and Tyr120, they may contribute to binding entropy. On the other hand, distribution between two different modes of binding in the crystal and the liquid phase might be different, so our prediction based on entropy calculation should be verified with nuclear magnetic resonance (NMR), rather X-ray structures. Finally there is a possibility that a higher resolution X-ray structure determination of the protein–ligand complexes discussed here will show a complete agreement with our entropy calculations.

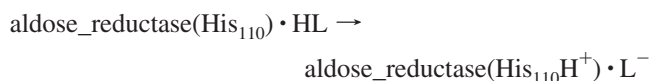
Renin–Diaminopyrimidines. Calculations were done for three ligands (compounds I–III), assuming single binding modes determined by the ligand coordinates in the corresponding PDB structures 2iko, 2iku, and 2il2.

Data in Table 4 show that, given the experimental uncertainty of 4 kJ/mol, predicted values of $-T\Delta S$ for compounds I and II are in fair agreement with the experimental results of Sarver et al.⁵⁰ For compound II, experiment shows the lack of entropy penalty (actually a small favorable effect) upon binding. Our calculations showing the entropy penalty of 7.5 kJ/mol does not reproduce this finding. It is likely that the loss of ligand entropy when the ligand binds to renin is compensated for the simultaneous increase in protein entropy, and as mentioned above, our current model does not include such effects.

Aldose Reductase–Sorbinil/Fidarestat. The only difference between sorbinil (compound IV) and the drug fidarestat (compound V) is the amide group missing in the former. Its oxygen forms a H-bond with the backbone NH group of Leu300 in the human enzyme. Fidarestat is bound tighter than sorbinil, which makes the former an efficient inhibitor of aldose reductase, able to slow the progression of diabetic neuropathy.⁵¹ Microcalorimetry results of Petrova et al.⁵² reported in Table 4 suggest however that the entropy penalty upon binding is twice as big as in the case of sorbinil. Our calculations based on the assumption of single binding modes for both compounds given in the crystal structures 1pwm (for fidarestat) and 2pdk (for sorbinil) do not confirm that finding. Predicted values of binding entropy $-T\Delta S$ are

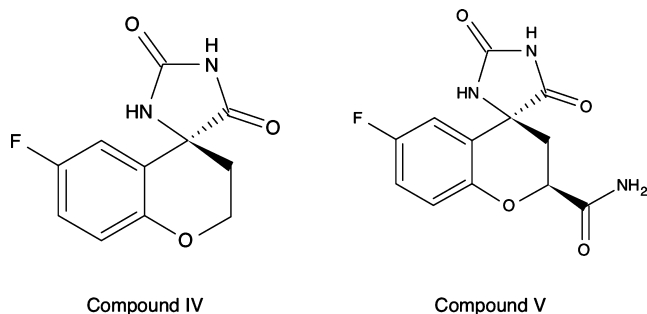


comparable for both compounds and agree well with the experimental value for sorbinil. More recent, very high-resolution (0.78 Å) crystallography data of Zhao et al.⁵³ offer a likely explanation of the discrepancy. It has been found that the bound ligand is deprotonated at the imide nitrogen and that the adjacent catalytic His110 is protonated. In other words, a strong electrostatic cation–anion attraction not only contributes to the large protein–ligand interaction (measured ΔH of binding is -75.5 kJ/mol⁵²) but also decreases the entropy of both the ligand and the side chain of His110 by restricting their motion. Two possible mechanisms could lead to the formation of the observed salt bridge: first, Zhao et al.⁵³ suggest a two step process in which the binding of a neutral ligand is followed by a proton-transfer reaction, and second, a direct binding of an anionic form of fidarestat by the enzyme with a protonated His110. The second mechanism cannot be ruled out because the estimated fidarestat pK_a is in the range of 7.9–8.5 pH units (estimation was made using the data for phenytoin and analogs),⁵⁴ which indicates that at the pH = 8, at which the ITC experiments were done,⁵² the fraction of ionized fidarestat in solution is at least 30%. We used our entropy calculation method to evaluate the binding entropy of the anionic form of fidarestat according to the above two scenarios. First, we calculated the entropy of a deprotonated fidarestat in the active site of aldose reductase with positively charged His110. The result of this calculation enabled us to estimate the entropy change for a proton-transfer reaction:

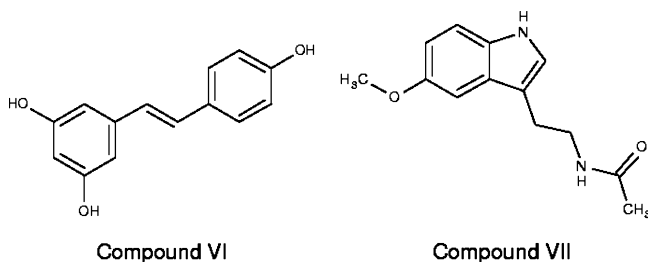


where HL is a neutral form of fidarestat. The estimated value of $T\Delta S$ at room temperature for the above reaction is -6.7 kJ/mol, so the overall calculated entropy change upon binding is $(-13.1-6.7) = -19.8$ kJ/mol, much closer to the experimental value of -28.8 kJ/mol. Second, the entropy change for direct binding of anionic forms of both fidarestat and sorbinil were calculated, and the results shown in Table 4. It is seen that the calculated values for the anionic form of fidarestat are essentially identical for both mechanisms,

so our method does not discriminate between the two possibilities. For both mechanisms of fidarestat anion binding, the calculated values are significantly more negative than in the case of the neutral ligand. Still more negative by about 9 kJ/mol, the experimental value probably reflects the restriction of protonated His110 motion in the electrostatic field of the negatively charged ligand, but as we mentioned earlier, our current model is not able to capture any protein entropy change contribution. The result for sorbinil suggests that the neutral form of the latter is bound by aldose reductase.



Quinone Reductase 2–Resveratrol/Melatonin. Structures of resveratrol, a natural polyphenol found in wine, and neurohormone melatonin are depicted as compounds VI and VII. The molecules bind in such a way that their aromatic rings are positioned in parallel to the isoalloxazine ring of a cofactor FAD. Initial positions of both ligands and the corresponding QR2 protein coordinates were taken from the PDB entries 1sg0 and 2qx4, respectively. Calculated and experimental values derived from ITC experiments of Calamini et al.⁵⁵ are shown in Table 4. Calculated values of $-T\Delta S$ for both resveratrol and melatonin are in very good agreement with experiment. Although given the approximate model used in our calculations, the ideal agreement for melatonin seems to be fortuitous.



Binding Entropy, Summary. Figure 7 shows the calculated binding entropies vs experimental ITC data for all tested protein–ligands complexes. The squared correlation coefficient ($R^2 = 0.81$) is relatively high, however, its confidence limits are large. In addition, considering that the calculated correlation is based only on those binding modes for hexanol and heptanol, which produce the closest agreement with experiment and preselection of those protein–ligand complexes for which protein contribution to the binding entropy is presumably negligible or low, at this moment, we are not able to evaluate the reliability of the method for binding entropy prediction. Instead we are considering these preliminary results as encouraging for further improvement of a high-throughput method for binding entropy prediction. Data for aldose reductase/fidarestat show the importance of both ligand and protein charge states,

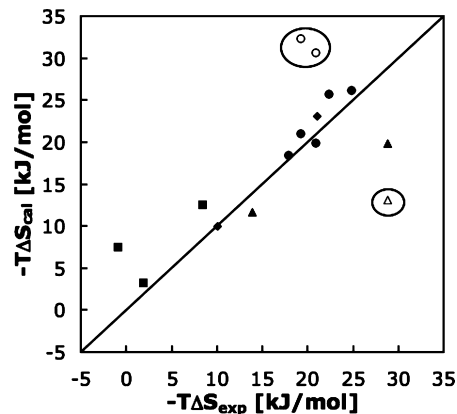


Figure 7. Calculated vs experimental binding entropies for protein–ligand systems selected for current study: (MUP-I)–*n*-alcohols (●), renin–diaminopyrimidines (■), aldose reductase–sorbinil/fidarestat (▲), and quinone reductase–resveratrol/melatonin (◆). Straight line represent ideal agreement. $R^2 = 0.81$, and its 95% confidence interval is (0.45–0.94). Points in ovals represent calculated data for hexanol and heptanol in binding mode I (structure I) and neutral form of fidarestat, respectively, and are not included to correlation.

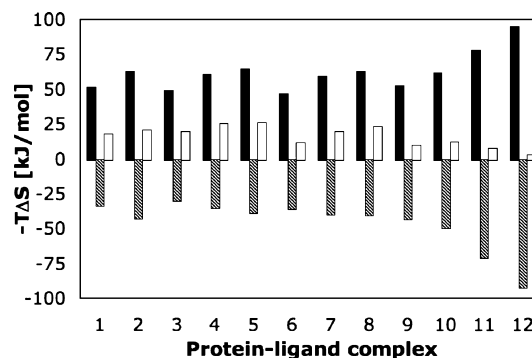


Figure 8. Calculated configurational entropy change (solid bars) and total solvation entropy change (shaded bars) upon binding. Empty bars show the $-T\Delta S_{\text{bind}}$ values from Figure 7. Numbers on horizontal axis correspond to protein–ligand complexes: MUP-I–pentanol (1), MUP-I–hexanol (2), MUP-I–heptanol (3), MUP-I–octanol (4), MUP-I–nonanol (5), aldose reductase–sorbinil (6), aldose reductase–fidarestat (7), quinone reductase–resveratrol (8), quinone reductase–melatonin (9), renin–compound I (10), renin–compound II (11), and renin–compound III (12).

while data for MUP-I/hexanol and MUP-I/heptanol show the necessity of considering multiple binding modes.

Given the correlation shown in Figure 7, it is tempting to search for a dominant calculated entropy term standing behind it. Figure 8 shows the configurational and the total solvation entropy change upon binding for the same 12 protein–ligand complexes which produced the correlation in Figure 7.

It is seen that both configurational and solvational entropy components play equally important roles in the binding entropy, so neither configurational entropy nor solvation entropy change alone is responsible for the correlation shown in Figure 7. In some protein–ligand complexes (like in the case of MUP-I complexed with alcohols), the loss of the configurational entropy is significantly larger than the entropy gain resulting from the solvation entropy change, while in

other cases (renin complexed with diaminopyrimidines), both components largely compensate for each other which results in little or no entropy penalty upon binding.

The ultimate goal of this work is to provide an accurate estimate of the entropy of binding of a small, drug-like molecule to a protein. Such an estimate is either time-consuming and difficult to calculate, for instance requiring the extensive sampling available from molecular dynamics, or is provided by very crude methods, such as penalty terms from the number of rotatable bonds. What we attempt here is to explore whether methods of intermediate speed and complexity can none the less be accurate. If so, it will provide a significant component of the long sought for computational approach to affinity prediction and its application to drug discovery.

Conclusions

- The Hessian matrix generated by the quasi-Newton optimizers of molecular structure with the use of a good quality force field can be used to predict vibrational entropy in the gas-phase, solution, and protein receptor environments.
- The simplified analytical solvation model formulated recently by Grant et al.¹⁸ can be used for ligand entropy calculations in solution.
- Entropy of molecules containing rotatable bonds is underestimated probably due to anharmonicity of the low-frequency torsional movements.
- The use of scaled particle theory (SPT) for the determination of the total solution entropies of compounds is promising. It eliminates the uncertainty associated with the value of surface tension coefficient used in the surface area (SA) model of hydrophobic solvation.
- Further work is required to address the anharmonicity of the quasi-Newton frequencies, the still unsatisfactory hydrophobic solvation model, and the contribution of protein receptors structural changes to the ligand binding entropy changes. This includes the usage of more accurate potential (force field, atomic charges on protein receptor, and analysis of protein ionization states), and entropic effects due to hydrogen bond and nonelectrostatic (van der Waals) solute-solvent interactions. Estimation of protein entropy with the use of a quasi-Newton Hessian matrix will require the usage of more efficient optimization techniques, such as limited memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) technique.
- Our results for aldose reductase-fidarestat system strongly suggest that pK_a analysis of both ligand and protein receptor should be included in the entropy estimation for protein-ligand binding.

Appendix

List of molecules which in addition to those listed in the caption for Figure 1 were used to determine the frequency scaling factor in entropy calculations.

Hydrocarbons.

Ethane, propane, *n*-butane, *n*-pentane, *n*-hexane, *n*-heptane, *n*-octane, *n*-nonane, *n*-decane, 2-methylpropane, 2-methylbutane, 2-methylpentane, 2-methylhexane, 3-methylhexane, propylene, 2,2-dimethylpropane, 2,3-dimethylbutane, 2,2-dimethylbutane, 1-butene, 1-pentene, *trans*-2-butene, *cis*-2-butene, *trans*-2-pentene, *cis*-2-pentene, 1,2-butadiene, 1,3-butadiene, propyne, 1-butyne, 2-butyne, butadiyne, biphenyl, 1,2-dimethylbenzene, 1,2,3-trimethylbenzene, 1,2,4-trimethylbenzene, pentamethylbenzene, and hexamethylbenzene.

Oxygen Compounds.

Ethanol, 2-propanol, *n*-propanol, *n*-butanol, *tert*-butyl alcohol, cyclohexanol, allyl alcohol, 2-butanol, *o*-cresol, *m*-cresol, *p*-cresol, dimethyl ether, diethyl ether, di-*n*-propyl ether, di-*n*-butyl ether, methyl-ethyl ether, methyl-propyl ether, tetrahydrofuran, 1,4-dioxan, acetaldehyde, propanal, acetone, butanal, pentanal, methyl-ethyl ketone, methyl-propyl ketone, diethyl ketone, cyclohexanone, acetic acid, methyl formate, and ethyl acetate.

Sulfur Compounds.

Ethanethiol, 1-propanethiol, 1-butanethiol, 1-pentanethiol, 1-hexanethiol, 1-heptanethiol, 1-octanethiol, 1-nonanethiol, 1-decanethiol, 2-propanethiol, 2-butanethiol, cyclopentanethiol, 2-methyl-1-propanethiol, 2-methyl-2-propanethiol, 2-methyl-2-butanethiol, benzenethiol, diethyl sulfide, dimethyl sulfide, ethyl-methyl sulfide, isopropyl-methyl sulfide, propyl-methyl sulfide, butyl-methyl sulfide, propyl-ethyl sulfide, butyl-ethyl sulfide, diisopropyl sulfide, pentyl-methyl sulfide, dipropyl sulfide, butyl-propyl sulfide, pentyl-ethyl sulfide, hexyl-methyl sulfide, dibutyl-sulfide, hexyl-ethyl sulfide, heptyl-methyl sulfide, dipentyl sulfide, *t*-butyl-methyl sulfide, diethyl-disulfide, dimethyl-disulfide, dipropyl-disulfide, dibutyl-disulfide, dimethyl-sulfoxide, dimethyl-sulfone, thiacyclobutane, thiacyclopentane, thiacyclohexane, thiacycloheptane, 2-methylthio-phene, and 3-methylthiophene.

Nitrogen Compounds.

Methylamine, ethylamine, *n*-propylamine, *n*-butylamine, *n*-pentylamine, *n*-hexylamine, ethylenediamine, 2-aminobutan, *t*-butylamine, dimethylamine, diethylamine, trimethylamine, triethylamine, aniline, propionitrile, butyronitrile, acrylonitrile, hydrazine, pyrrolidine, 2-picoline, nitroethane, nitropropane, nitrobutane, 2-nitropropane, 2-nitrobutane, ethyl nitrate, *n*-propyl-nitrate, and isopropyl-nitrate.

Halogen Compounds.

Fluoroethane, 1-fluoropropane, 1,1-difluoroethane, 1,1,1-trifluoroethane, hexafluoroethane, 1-chloropropane, chloroethane, 1-chlorobutane, 1-chloropentane, 2-chloropropane, 2-chlorobutane, 1-chloro-3-methylbutane, 1-chloro-2-methylpropane, 2-chloro-2-methylpropane, 1,2-dichloropropane, 1,2-dichloroethane, 2-chloro-2-methylbutane, 1,3-dichloropropane, 1,1-dichloroethane,

1,1,2-trichloroethane, 2,2-dichloropropane, 1,2,3-trichloropropane, 1,1,2,2-tetrachloroethane, 3-chloro-1-propene, pentachloroethane, hexachloroethane, acetylchloride, bromoethane, 1-bromopropane, 1-bromobutane, 1-bromopentane, 2-bromopropane, 2-bromobutane, 2-bromo-2-methylpropane, 1,2-dibromoethane, 1,2-dibromopropane, 1,2-dibromobutane, 2,3-dibromobutane, 2,3-dibromo-2-methylbutane, 3-bromo-1-propene, 1-bromopropyne, iodoethane, 1-iodopropane, 2-iodo-2-methylpropane, 2-iodopropane, 1,2-diiodoethane, 1,2-diiodopropane, 1,2-diiodobutane, and 3-iodo-1-propene.

References

- (1) McQuarie, D. *Statistical Mechanics*; Harper & Row: New York, 1976; pp 129–149.
- (2) Edholm, O.; Berendsen, H. *Mol. Phys.* **1984**, *51*, 1011–1028.
- (3) Karplus, M.; Kushick, J. *Macromolecules* **1981**, *14*, 325–332.
- (4) Andricioaei, I.; Karplus, M. *J. Chem. Phys.* **2001**, *115*, 6289–6292.
- (5) Schlitter, J. *Chem. Phys. Letters* **2003**, *215*, 617–621.
- (6) Mammen, M.; Shakhnovitch, E.; Deutch, J.; Whitesides, G. *J. Org. Chem.* **1998**, *63*, 3821–3830.
- (7) Amzel, L. *Proteins* **1997**, *28*, 144–149.
- (8) Siebert, X.; Amzel, L. *Proteins* **2004**, *54*, 104–115.
- (9) Graziano, G. *J. Phys. Chem. B* **2005**, *109*, 12160–12166.
- (10) Chang, C.; Gilson, M. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.
- (11) Chen, W.; Chang, C.; Gilson, M. *Biophys. J.* **2004**, *87*, 3035–3049.
- (12) Chang, C.; Chen, W.; Gilson, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534–1539.
- (13) Halgren, A. *J. Comput. Chem.* **1996**, *17*, 490–519, 520–552, 553–586, 587–615, 616–641.
- (14) *Omega2.1*, OpenEye Scientific Software, Inc.: Santa Fe, NM, 2006.
- (15) Boström, J. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137–1152.
- (16) Wlodek, S.; Skillman, A.; Nicholls, A. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 741–749.
- (17) Nocedal, J.; Wright, S. *Numerical Optimization*; Glynn, P., Robinson, S. M., Eds.; Springer-Verlag: New York, 1999.
- (18) Grant, J.; Pickup, B.; Sykes, M.; Kitchen, C.; Nicholls, A. *Chem. Phys. Lett.* **2007**, *441*, 163–166.
- (19) *CRC Handbook of Chemistry and Physics*; Lide, D., Ed.; CRC Press: New York, 1999.
- (20) Suresh, S.; Naik, V. *J. Chem. Phys.* **2000**, *113*, 9727–9732.
- (21) Reynolds, J.; Gilbert, D.; Tanford, C. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 2925–2927.
- (22) Sitkoff, D.; Sharp, K.; Honig, B. *Biophys. Chem* **1994**, *51*, 397–409.
- (23) Hermann, R. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 4144–4145.
- (24) Sharp, K.; Nicholls, A.; Fine, R.; Honig, B. *Science* **1991**, *252*, 106–109.
- (25) Chandler, D. *Nature* **2005**, *437*, 640–647.
- (26) Pierotti, R. *Chem. Rev.* **1976**, *76*, 717–726.
- (27) Graziano, G. *J. Chem. Soc., Faraday Trans.* **1998**, *94*, 3345–3352.
- (28) Head-Gordon, T.; Hura, G. *Chem. Rev.* **2002**, *102*, 2651–2670.
- (29) Grant, J.; Gallardo, J.; Pickup, B. *J. Comput. Chem.* **1996**, *17*, 1653–1666.
- (30) Wilhelm, E.; Batino, R. *J. Chem. Phys.* **1971**, *55*, 4012–4017.
- (31) Mouritz, F.; Rummens, F. *Can. J. Chem.* **1977**, *55*, 3007–3020.
- (32) Cuadros, F.; Mulero, A.; Cachadina, I. *Int. Rev. Phys. Chem.* **1995**, *14*, 205–213.
- (33) Cachadina, I.; Mulero, A. *J. Phys. Chem. Ref. Data* **2007**, *36*, 1133–1139.
- (34) Wang, J.; Xie, X.; Hou, T.; Xu, X. *J. Phys. Chem. A* **2007**, *111*, 4443–4448.
- (35) *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Linstrom, P.; Mallard, W., Eds.; National Institute of Standards and Technology: Gaithersburg, MD; <http://webbook.nist.gov>. Accessed January 30, 2010.
- (36) Domalski, E.; Hearing, E. *J. Phys. Chem. Ref. Data* **1993**, *22*, 805–1159.
- (37) Blom, C.; Altona, C. *Mol. Phys.* **1976**, *31*, 1377–1391.
- (38) Rauhut, G.; Pulay, P. *J. Phys. Chem.* **1995**, *99*, 3093–3100.
- (39) Barone, V. *J. Chem. Phys.* **2004**, *120*, 3059–3065.
- (40) Barone, V. *J. Chem. Phys.* **2005**, *122*, 014108-1–014108-10.
- (41) Plyasunova, N.; Plyasunov, A.; Shock, E. *Int. J. Thermophys.* **2004**, *25*, 351–360.
- (42) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563–595.
- (43) Boyer, J.; Francis, M.; Boerio-Goates, J. *J. Chem. Thermodyn.* **2003**, *35*, 1917–1928.
- (44) Tewari, Y.; Gery, P.; Vaudin, M.; Mighell, A.; Klein, R.; Goldberg, R. *J. Chem. Thermodyn.* **2004**, *36*, 645–658.
- (45) Lee, B. *Biopolymers* **1991**, *31*, 993–1008.
- (46) Graziano, G. *Biophys. Chem.* **2003**, *104*, 393–405.
- (47) Irida, M.; Nagayama, K.; Hirata, F. *Chem. Phys. Lett.* **1993**, *207*, 430–435.
- (48) Berman, H.; Henrick, K.; Nakamura, H. *Nat. Struct. Biol.* **2003**, *10*, 980–980.
- (49) Malham, R.; Johnstone, S.; Bingham, R.; Barratt, E.; Phillips, S.; Laughton, C.; Homans, S. *J. Am. Chem. Soc.* **2005**, *127*, 17061–17067.
- (50) Sarver, R. *J. Anal. Biochem.* **2007**, *360*, 30–40.
- (51) Hotta, N.; Toyota, T.; Matsuoka, K.; Shigeta, Y.; Kikkawa, R.; Kaneko, T.; Takahashi, A. *Diabetes Care* **2001**, *24*, 1776–1782.
- (52) Petrova, T.; Steuber, H.; Hazemann, I.; Cousido-Siah, A.; Mitschler, A.; R.Chung.; Oka, M.; Klebe, G.; El-Kabbani, O.; Joachimiak, A.; Podjarny, A. *J. Med. Chem.* **2005**, *48*, 5659–5665.
- (53) Zhao, H.; Hazemann, I.; Mitschler, A.; Carbone, V.; Joachimiak, A.; Ginell, S.; Podjarny, A.; El-Kabbani, O. *J. Med. Chem.* **2008**, *51*, 1478–1481.
- (54) Porter, R.; Meldrum, B. *Basic and Clinical Pharmacology*, 5th ed.; Appleton & Lange: Norwalk, CT, 1992.
- (55) Calamini, B.; Santarsiero, B.; Boutin, J.; Mesecar, A. *Biochem. J.* **2008**, *413*, 81–91.

A New Empirical Correction to the AM1 Method for Macromolecular Complexes

Michael E. Foster and Karl Sohlberg*

*Department of Chemistry, Drexel University, 3141 Chestnut Street,
Philadelphia, Pennsylvania 19104*

Received April 2, 2010

Abstract: Modeling systems that are governed by van der Waals (dispersion) interactions using empirically corrected DFT methods is becoming increasingly popular due to the promise of a CCSD(T) level accuracy at the computational cost of DFT. Although, DFT methods are computationally efficient in comparison to the CCSD(T) method, currently, structural optimizations using DFT methods are generally only feasible for systems of less than a few hundred atoms. We seek a method applicable to macromolecular complexes. In order to model such large systems, empirically corrected semiempirical methods appear to be an attractive alternative. As with most common DFT methods, the popular semiempirical methods (e.g., AM1) also do not model long-range dispersion (and therefore an empirical correction term is desirable), but this is not their only shortcoming. For weakly interacting systems, hydrogen bonding also poses a concern. A new empirically corrected AM1 method that uses two empirical correction terms, one for dispersion and one for hydrogen bonding interactions, is presented and termed AM1-FS1. This new empirically corrected AM1 method has been parametrized to a diverse training set of 66 complexes that includes nonequilibrium structures and yields sub-kilocalorie accuracy in the prediction of intermolecular interaction energies. More significantly, AM1-FS1 achieves this result with substantially less parametrization than existing empirically corrected semiempirical methods and *without modification of the original AM1 parameters* so that it retains both the computational efficiency and predictive power for thermo-chemical quantities of the original AM1 Hamiltonian. The performance of AM1-FS1 is also tested on several carbon nanostructure complexes and pseudorotaxanes and is found to produce results in very good agreement with the best first-principles calculations.

1. Introduction

Accurately and efficiently modeling nonbonding interactions, especially van der Waals interactions and hydrogen bonding, is a difficult task, but essential for the correct description of many systems of chemical and biological importance, such as the structure of molecular crystals,^{1,2} the conformational preference^{3–9} and folding of proteins, and the stability of two strands of DNA in a double helix.^{3,10} These types of interactions are also important in macromolecular host/guest chemistry,^{11–14} which motivates this work.

Density functional theory (DFT) is the most widely used quantum mechanical (QM) technique for chemical calculations due in part to its ability to accurately describe chemical and physical properties for a diversity of systems, often at modest computational expense. A major shortcoming with DFT is the inability of most popular XC functionals (exchange-correlation functionals) to accurately model long-range van der Waals (dispersion) interactions. Therefore, these methods predict systems like the benzene dimer to be unbound (Figure 1). Currently, an increasingly popular approach to overcome this hurdle is to add an empirical correction to the DFT total energy. Empirically corrected DFT methods for dispersion interactions, coined DFT-D, have become popular due to

* Corresponding author e-mail: kws24@drexel.edu.

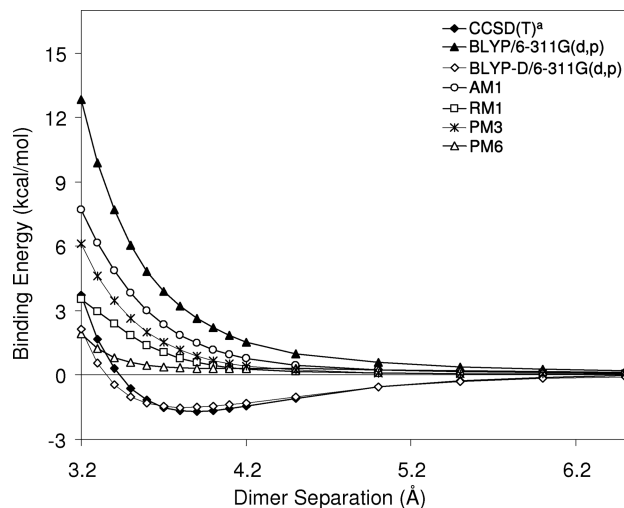


Figure 1. Potential energy curves for the parallel benzene dimer determined with various quantum mechanical methods. Superscript a refers to ref 23.

their success with essentially no added computational expense. Figure 1 shows the dramatic improvement that is achieved by adding an empirical correction term. Not only is the complex now predicted to be bound but excellent agreement with CCSD(T) results is achieved, the current “gold standard” in computational chemistry. For relatively small systems, DFT-D methods are computationally feasible and should provide quite accurate results, but modeling macromolecular host/guest complexes can be extremely computationally expensive. Therefore, alternative methods need to be explored.

Semiempirical (SE) techniques such as AM1,¹⁵ PM3,¹⁶ RM1,¹⁷ and PM6¹⁸ are sufficiently computationally efficient for modeling systems composed of hundreds or even thousands of atoms but typically perform poorly for dispersion and hydrogen bonding. These semiempirical methods are essentially incapable of modeling dispersion bound complexes because the form of the semiempirical wave function completely neglects electron correlation. Even qualitatively reliable modeling of dispersion-bound macromolecular systems, such as complexes of carbon nanostructures, is therefore out of the question; SE methods predict such complexes to be unbound. That is, the wrong sign of the interaction is predicted. Again, considering the parallel benzene dimer (Figure 1), which is a classic test case used for predicting the likely accuracy of a method for modeling complexes of carbon nanostructures, we see that the SE methods fail.

The accuracy of SE methods in modeling dispersion-bound systems can be dramatically improved by adding an empirical correction term. McNamara and Hillier¹⁹ reported adding an empirical correction term to the AM1 and PM3 methods to incorporate dispersion interactions but found the overall results to be unsatisfactory. Therefore, to gain further improvements, in particular to improve the accuracy with which hydrogen-bonding is modeled, they reoptimized 18 of the original AM1 parameters, using the S22 database²⁰ as a training set. The resulting method, with *both* reoptimized semiempirical parameters *and* an empirical correction term,

is referred to as AM1-D. (They have also produced an analogous PM3-D.) These empirically corrected methods show a substantial improvement in accuracy (over the corresponding original SE methods) for predicting intermolecular interaction energies, but at a significant cost. As detailed below, AM1-D is nearly 25-fold less accurate in the prediction of heats of formation than the original AM1 method. We seek a method that leads to good accuracy in the prediction of structures and interaction energies for macromolecular complexes, without sacrificing predictive power for the heat of formation.

More recently Řezáč and colleagues²¹ published an empirically corrected PM6 method for modeling dispersion and hydrogen-bonding interactions, named PM6-DH. To incorporate dispersion interactions, the group used the empirical correction described by Jurecka et al.,²² which is also discussed here in section 2.1. To improve the PM6 method for H-bonding, they included a second correction term involving three parameters. The correction term is applied to H-bonding situations, but not all types of H bonds are modeled with the same term. The group identified eight types of H bonds and used a different set of three parameters for each type, for a total of 24 H-bonding parameters. In their defense, it should be noted that they did use a relatively large training set to determine these H-bonding parameters. The major shortcoming of this method, as they acknowledge, is that knowledge of atom connectivity is required. One of the major benefits of QM techniques is that atom connectivity is not required, allowing bond formation/deformation to be modeled. Although, Řezáč and colleagues have obtained good results, for a method to be widely used, atom conductivity information should not be required. In addition to the limitations introduced, input of atom connectivity information is sufficiently burdensome to deter routine use, especially for macromolecular complexes where there may be thousands of atom–atom interactions that must be distinguished.

Herein, we present an empirical correction for the AM1 method that is suitable for modeling macromolecular complexes and avoids the above-noted shortcomings of existing techniques. We have chosen to apply separate empirical correction terms for dispersion and hydrogen bonding. Our method requires significantly less parametrization than the AM1-D and PM3-D methods of McNamara and Hillier¹⁹ and also the PM6-DH method of Řezáč et al.²¹ Additionally, it is important to note that we have not altered any of the original AM1 parameters. Such changes can have deleterious effects on predictions of properties not based strictly on the total energy or its derivatives, such as heats of formation, ionization potentials, and dipole moments, if these quantities are not taken into consideration during reparameterization. Our method also does not require knowledge of atom connectivity. We will henceforth refer to our new method as “AM1-FS1”. AM1-FS1 achieves results that are comparable to (and in many cases better than) those of other empirically corrected SE methods, with *significantly* less parametrization and with no reparameterization of the AM1 method. The main objective of AM1-FS1 is to accurately model macromolecular host/guest systems that are currently

out of reach of DFT-D techniques. AM1-FS1 aims to not only accurately predict energies but also reasonable structures upon geometry optimization, since structural optimization is one of the main uses for such a technique. Herein, the accuracy of AM1-FS1 is tested by comparing interaction energies and distances to CCSD(T) and SAPT results; comparisons are also made with other empirically corrected semiempirical techniques.

2. Theory

2.1. Dispersion Correction. To correct the AM1 method for dispersion interactions, we have employed a method used by Grimme²⁴ with a slight modification suggested by Jurečka et al.²² The resulting dispersion correction is of the form

$$E_{\text{dis}} = -\frac{C_6^{ij}}{r_{ij}^6} f_{\text{damp}}(r_{ij}) \quad (1)$$

where r_{ij} is the atom–atom separation, C_6^{ij} is the dispersion coefficient, and f_{damp} is a damping function of the form

$$f_{\text{damp}}(r_{ij}) = \frac{1}{1 + \exp(-d(\frac{r_{ij}}{S_R R_{\text{vdw}}} - 1))} \quad (2)$$

This damping function depends on the equilibrium van der Waals separation (R_{vdw}) and the pairwise atom separation (r_{ij}). The damping function also depends on two unitless parameters, S_R and d , which have been optimized to a training set as discussed at length in section 2.3. The damping function operates as a switching function, turning off the dispersion term at short range. This is required because the SE wave function already models short-range repulsive interactions. Thus, the popular 6–12 Lennard-Jones (LJ) potential is not suitable for use as a dispersion correction since a repulsive term is involved (see ref 25 for a more detailed discussion and graphical representations).

It should be noted that we tried employing the global scaling factor, used by Grimme,^{24,26} instead of scaling the equilibrium van der Waals separation (S_R); however, a smaller root-mean-square error (RMSE) was obtained on our training set using S_R (discussed in section 2.3). Scaling R_{vdw} seems theoretically well motivated, since this allows only the short-range interactions to be tailored and leaves untouched the long-range interactions for which the correct functional form of the interaction is known to follow r^{-6} (see ref 25 for a more detailed discussion and graphical representations).

Another decision concerns the choice of combination rules used for obtaining C_6^{ij} and R_{vdw} . We have chosen to employ the geometric mean and simple average combination rules for determining C_6^{ij} and R_{vdw} , respectively:

$$C_6^{ij} = \sqrt{C_6^i C_6^j}, R_{\text{vdw}} = \frac{R_i + R_j}{2} \quad (3)$$

The dispersion coefficients (C_6^i and C_6^j) and van der Waals radii (R_i and R_j) for the different atoms were obtained from Grimme's 2006 publication.²⁴ The decision of using these particular combination rules was not made without consider-

ing other options. For C_6^{ij} , both the harmonic mean²⁶ and the combination rule suggested by Wu and Yang,²⁷ which uses the Slater–Kirkwood effective number of electrons, were considered. For R_{vdw} , the cubic mean suggested by Halgren²⁸ was also considered. We have also considered all possible combinations and found that the parameters (S_R and d) seemed to adjust to accommodate the different combination rules. The combination rules employed yielded the lowest RMSE for our training set. It should be noted that only Grimme's 2006 published dispersion coefficients and van der Waals radii values were considered. This dispersion correction scheme to the AM1 method has been implemented into a locally modified version of GAMESS.²⁹

2.2. Hydrogen-Bonding Correction. Correcting the AM1 method for hydrogen bonding is a more difficult task than correcting for its neglect of dispersion since hydrogen-bonding interactions are already in part considered, given their partial electrostatic nature. It can be seen by looking at the H-bonded systems (1–7) in the S22 database (Table 1) that the AM1 method severely underbinds such complexes. The AM1 method does, however, produce more reasonable interaction energies for hydrogen-bonded systems upon geometry optimizations (Table 2); this is because AM1 generally predicts dispersion-bound complexes to be unbound, while for H-bonded complexes it predicts some binding, but generally with an unphysically large equilibrium separation (see Table 3). Thus, to improve the AM1 method for predicting H-bonding systems, the strength of these interactions needs to be increased at medium-to-short range. We have achieved this by adding a post-SCF pseudoelectrostatic term of the form

$$E_{\text{HB}} = \alpha_1 \frac{Q_i Q_j}{r_{ij}} \cos^2(\theta) f_{\text{damp}2}(r_{ij}) \quad (4)$$

where α_1 is a global scaling factor, Q_i and Q_j are the AM1 Coulson charges³⁰ (which are referred to as MOPAC charges in GAMESS²⁹), r_{ij} is the H---Y separation, θ is the XH---Y angle, and $f_{\text{damp}2}$ is a damping function of the form

$$f_{\text{damp}2}(r_{ij}) = \exp[-(r_{ij} - \alpha_2 R_{\text{vdw}})^2 / \alpha_3 (1 + \alpha_4 (r_{ij} - \alpha_2 R_{\text{vdw}}))^2] \quad (5)$$

where α_2 , α_3 , and α_4 are parameters and all other terms have the same meanings as in the dispersion correction. In this case, however, R_{vdw} is defined as the cubic mean

$$R_{\text{vdw}} = \frac{R_i^3 + R_j^3}{R_i^2 + R_j^2} \quad (6)$$

The cubic mean is used in this case because it yields a slightly smaller RMSE for the F66 training set than using the simple average combination rule.

The damping function is an asymmetric distribution function (see Figure 2A) that turns the hydrogen-bonding function on or off over an appropriate range for correcting the AM1 method. To achieve an asymmetric distribution, three parameters (α_2 , α_3 , α_4) have been introduced, giving a total of four parameters in the H-bonding correction. We have optimized these four parameters to improve upon H

Table 1. Single-Point Interaction Energies (kcal/mol) at the S22 Geometries^a

no.	molecule (symmetry)	ref values	AM1	PM3	AM1-D	PM3-D	PM6-DH	AM1-FS1
Hydrogen-Bonded Complexes								
1	(NH ₃) ₂ (C2h)	-3.17	-0.78	0.77	-3.43	-1.77	-3.74	-1.60
2	(H ₂ O) ₂ (Cs)	-5.02	-2.89	-2.79	-7.29	-5.14	-4.67	-5.53
3	formic acid dimer (C2h)	-18.61	1.54	-9.91	-15.45	-18.57	-17.39	-16.06
4	formamide dimer (C2h)	-15.96	-12.02	-8.08	-17.16	-15.37	-15.39	-15.75
5	uracil dimer (C2h)	-20.65	-5.79	-11.32	-20.15	-20.30	-18.84	-20.80
6	2-pyridoxine2-aminopyridine (C1)	-16.71	-4.45	-7.46	-16.50	-17.52	-17.35	-14.73
7	adenine thymine WC (C1)	-16.37	-4.28	-6.79	-16.58	-17.33	-17.83	-16.29
Complexes with Predominant Dispersion Contribution								
8	(CH ₄) ₂ (D3d)	-0.53	0.21	-0.25	-0.94	-1.24	-0.73	-0.61
9	(C ₂ H ₄) ₂ (D2d)	-1.51	-0.13	-1.11	-3.31	-3.60	-1.52	-2.27
10	benzene CH ₄ (C3)	-1.50	0.40	-0.19	-2.12	-2.42	-1.75	-1.79
11	benzene dimer (C2h)	-2.73	3.52	2.38	-2.90	-4.30	-3.62	-2.23
12	pyrazine dimer (Cs)	-4.42	2.49	3.90	-4.57	-4.20	-5.41	-3.81
13	uracil dimer (C2)	-10.12	0.12	5.80	-10.56	-6.78	-9.70	-8.47
14	indole benzene (C1)	-5.22	5.39	4.04	-4.04	-6.09	-5.20	-3.23
15	adenine thymine stack (C1)	-12.23	2.91	7.37	-12.20	-10.63	-12.78	-9.87
Mixed Complexes								
16	ethene ethine (C2v)	-1.53	-0.35	-0.82	-1.61	-1.85	-1.11	-1.36
17	benzene H ₂ O (Cs)	-3.28	-0.69	-1.47	-3.43	-3.65	-3.41	-2.78
18	benzene NH ₃ (Cs)	-2.35	-0.33	-0.59	-3.00	-2.96	-2.77	-2.65
19	benzene HCN (Cs)	-4.46	-0.81	-1.63	-4.44	-4.43	-3.20	-3.17
20	benzene dimer (C2v)	-2.74	0.37	-0.43	-3.85	-4.15	-2.84	-3.36
21	indole benzene T-shape (C1)	-5.73	-1.05	-1.25	-7.10	-6.65	-5.30	-4.63
22	phenol dimer (C1)	-7.05	-1.36	-1.37	-9.76	-7.52	-6.73	-6.91
	RMSE (hydrogen bonded)		11.64	7.77	1.56	0.76	1.07	1.37
	RMSE (dispersion bonded)		8.21	10.13	0.82	1.68	0.54	1.30
	RMSE (mixed bonded)		3.57	3.22	1.25	0.72	0.57	0.72
	RMSE		8.47	7.73	1.23	1.18	0.76	1.18
	MUE		6.54	5.94	0.85	0.90	0.59	0.88

^a The AM1-D and PM3-D results have been taken from ref 19, and the PM6-DH results from ref 21.

bonding for the AM1 method. A detailed discussion is presented in the next section.

The H-bonding correction function also depends on the square of the cosine of the XH---Y angle. This is motivated

Table 2. Geometry Optimized Energies (kcal/mol) for the Complexes in the S22 Database^a

no.	molecule (symmetry)	ref values	AM1	PM3	AM1-D	PM3-D	PM6-DH	AM1-FS1
Hydrogen-Bonded Complexes								
1	(NH ₃) ₂ (C2h)	-3.17	-1.39	-0.71	-3.03	-1.99	-3.92	-2.82
2	(H ₂ O) ₂ (Cs)	-5.02	-3.30	-3.55	-7.22	-6.53	-4.73	-5.59
3	formic acid dimer (C2h)	-18.61	-6.62	-9.58	-12.45	-16.16	-19.11	-17.76
4	formamide dimer (C2h)	-15.96	-2.06	-6.99	-14.64	-14.42	-15.01	-15.83
5	uracil dimer (C2h)	-20.65	-10.48	-10.70	-17.80	-18.83	-19.55	-25.06
6	2-pyridoxine2-aminopyridine (C1)	-16.71	-6.15	-7.06	-13.06	-18.32	-18.50	-15.16
7	adenine thymine WC (C1)	-16.37	-5.06	-6.90	-12.66	-18.66	-19.12	-21.10
Complexes with Predominant Dispersion Contribution								
8	(CH ₄) ₂ (D3d)	-0.53	-0.21	-0.32	-4.10	-2.38	-0.73	-2.46
9	(C ₂ H ₄) ₂ (D2d)	-1.51	-0.13	-1.08	-4.85	-4.11	-1.53	-4.09
10	benzene CH ₄ (C3)	-1.50	0.35	-0.20	-2.93	-2.88	-1.88	-2.84
11	benzene dimer (C2h)	-2.73	0.01	-0.02	-3.10	-4.59	-3.59	-2.21
12	pyrazine dimer (Cs)	-4.42	-0.34	-0.26	-4.87	-4.45	-5.74	-4.73
13	uracil dimer (C2)	-10.12	-6.05	-4.26	-11.25	-7.59	-10.03	-9.99
14	indole benzene (C1)	-5.22	-1.33	-1.65	-8.16	-6.26	-5.99	-6.51
15	adenine thymine stack (C1)	-12.23	-5.15	-6.50	-15.13	-11.70	-13.61	-12.59
Mixed Complexes								
16	ethene ethine (C2v)	-1.53	-0.57	-1.23	-2.47	-2.58	-1.17	-1.50
17	benzene H ₂ O (Cs)	-3.28	-1.03	-1.63	-3.90	-4.46	-3.95	-3.38
18	benzene NH ₃ (Cs)	-2.35	-0.80	-0.93	-4.04	-3.99	-3.82	-4.70
19	benzene HCN (Cs)	-4.46	-0.92	-1.85	-4.28	-4.40	-3.21	-2.46
20	benzene dimer (C2v)	-2.74	-0.09	-0.52	-4.22	-4.39	-2.85	-2.15
21	indole benzene T-shape (C1)	-5.73	-1.24	-1.67	-7.74	-7.20	-5.22	-5.88
22	phenol dimer (C1)	-7.05	-3.39	-4.33	-11.55	-8.95	-7.46	-8.87
	RMSE (hydrogen bonded)		9.90	8.04	3.38	1.82	1.40	2.55
	RMSE (dispersion bonded)		3.73	3.65	2.36	1.71	0.80	1.34
	RMSE (mixed bonded)		2.96	2.40	2.09	1.40	0.82	1.37
	RMSE		6.25	5.22	2.65	1.65	1.04	1.82
	MUE		4.82	4.09	2.16	1.51	0.82	1.28

^a The AM1-D and PM3-D results have been taken from ref 19 and the PM6-DH results from ref 21.

Table 3. Interaction Distances (Ångstroms) for the Complexes in the S22 Database^a

no.	molecule (symmetry)	ref values	AM1	PM3	AM1-D	PM3-D	AM1-FS1
Hydrogen-Bonded Complexes							
1	(NH ₃) ₂ (C2h)	2.504	2.784	3.241	2.646	2.726	2.668
2	(H ₂ O) ₂ (Cs)	1.952	2.094	1.809	1.911	1.769	1.932
3	formic acid dimer (C2h)	1.670	2.101	1.776	1.925	1.737	1.567
4	formamide dimer (C2h)	1.841	2.072	1.807	1.981	1.763	1.916
5	uracil dimer (C2h)	1.775	2.044	1.787	1.946	1.744	1.563
6	2-pyridoxine-2-aminopyridine (C1)	1.859, 1.874	2.511, 2.107	1.798, 1.815	1.980, 1.981	1.722, 1.768	1.760, 1.878
7	adenine thymine WC (C1)	1.819, 1.929	2.476, 2.101	1.780, 1.821	1.807, 2.018	1.708, 1.769	1.597, 1.893
Complexes with Predominant Dispersion Contribution							
8	(CH ₄) ₂ (D3d)	3.718	3.721	3.447	2.881	3.160	2.899
10	benzene CH ₄ (C3)	3.716	3.746	3.718	3.315	3.450	3.457
11	benzene dimer (C2h)	3.765	6.952	6.096	3.643	3.499	3.753
12	pyrazine dimer (Cs)	3.479	4.848	4.760	3.695	3.437	3.681
13	uracil dimer (C2)	3.166	5.805	6.732	3.097	3.406	3.007
14	indole benzene (C1)	3.498	5.572	5.520	4.448	3.415	4.378
15	adenine thymine stack (C1)	3.172	6.202	5.788	4.320	3.280	3.099
Mixed Complexes							
16	ethene ethine (C2v)	2.752	2.468	2.429	2.319	2.366	2.374
18	benzene NH ₃ (Cs)	3.592	4.092	4.025	2.995	3.069	3.014
19	benzene HCN (Cs)	3.387	3.472	3.694	3.228	3.343	3.303
20	benzene dimer (C2v)	3.513	5.225	3.606	3.253	3.370	3.351
21	indole benzene T-shape (C1)	3.210	3.811	3.807	3.010	3.233	3.208
22	phenol dimer (C1)	1.937, 4.921	2.174, 5.925	1.829, 5.712	2.001, 5.040	1.778, 5.265	2.016, 4.937
	RMSE (hydrogen bonded)		0.387	0.257	0.137	0.134	0.129
	RMSE (dispersion bonded)		2.015	1.962	0.644	0.272	0.473
	RMSE (mixed bonded)		0.929	0.598	0.336	0.315	0.301
	RMSE		1.277	1.171	0.419	0.249	0.326
	MUE		0.853	0.691	0.301	0.199	0.222

^a The AM1-D and PM3-D results have been taken from ref 19.

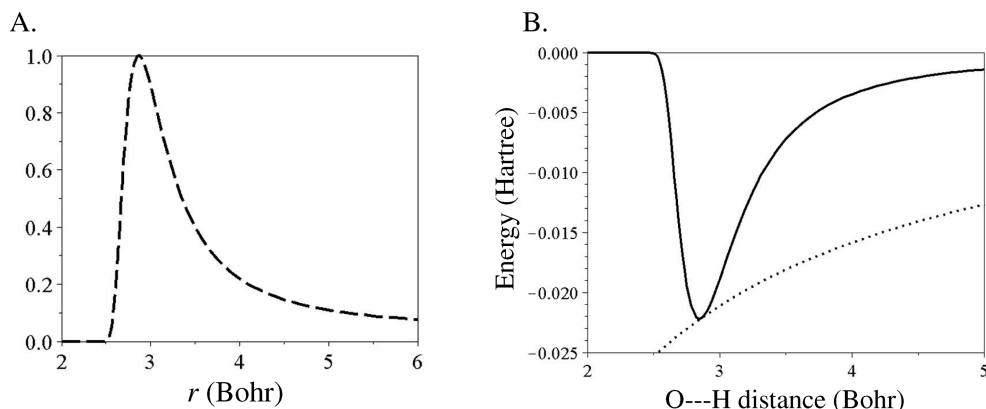


Figure 2. Graphical representation of the H-bonding damping function (A), the entire correction term (solid-line), and the electrostatic attractive portion (dotted-line) (B) used in the AM1-FS1 method. This model is for the case of the α -hydrogen atom (connected to the nitrogen atom) interacting with the parallel oxygen atom on the second monomer of the uracil dimer in the hydrogen bonding conformation. The MOPAC charges used correspond to the minimum energy structure (O---H, $R = 1.77$); this simplification has little effect on the functional form. This simplification has been used for graphical convenience.

by the observation that H-bonding interactions are directionally dependent.³¹ The cosine squared function was used instead of the cosine function because it approaches zero smoothly. We have also chosen to make the function zero for all angles less than 90° . This helps exclude cases that are not H-bonding, such as an α -H atom in a carboxylic acid interacting with the adjacent carboxylate O atom. By using an appropriate summing scheme, we are able to identify highly likely H-bonding scenarios without knowledge of atom connectivity. This is done by first identifying H atoms for which the nearest neighbor is an N, O, or F atom. These H atoms are then allowed to interact with other N, O, or F atoms.

The overall function (eq 4) is shown graphically in Figure 2B, where it can be seen that the function is only turned on over a short-range, peaking at approximately 2.8 bohr (1.5 Å; for the specific bonding scenario depicted). This is the behavior that is needed to improve the AM1 method for H-bonding, since these interactions only need to be increased over a short range and only at short distances. The nature of the charges (MOPAC) of the atoms involved in H-bonding, insures that eq 4 is negative, resulting in an attractive contribution. The charges are updated every optimization step. The optimization procedure also requires the gradient, which is determined by numerical differentiation. This correction

scheme to the AM1 method has been implemented into a locally modified version of GAMESS.²⁹

The hydrogen bonding correction scheme as described above is continuous for proton transfer under most conditions. In most cases, the correction term effectively turns off (i.e., is essentially equal to zero) before the proton reaches the halfway point in a proton transfer. For example, when a proton transfers between two formic acid molecules, the intercomponent oxygen–oxygen separation is about 2.7 Å; therefore, when the proton reaches the halfway point, the H–Y distance is about 1.35 Å (2.55 bohr) and the H-bonding correction term is approximately zero (see Figure 2B). The function is also continuous when the molecules are separated by a greater distance even though there is a nonzero correction at the halfway point because at this point the function is identical in both directions (H–Y equals X–H). When the proton passes the halfway point, the H-bonding correction term corrects in the opposite direction. If a proton is transferring across an asymmetric system, however, a discontinuity can occur since the charges on the X and Y atoms may not be the same. This discontinuity can be eliminated by evaluating eq 4 in both directions at all times and taking the correction to be a weighted sum of the two. This correction has been implemented into AM1-FS1 and has no effect on any of the binding energies reported in this manuscript, since evaluating the function in the opposite direction (considering the X–H bond to be H-bonding) leads to no correction because the X–H distance is essentially always less than 1 Å (1.9 bohr). In summary, the switching transition from one H-bonding situation to another is effectively continuous during a proton transfer. We currently cannot recommend AM1-FS1 for modeling proton transfers, since it has not been tested and more importantly because our training set does not contain data to parametrize for such situations. Nevertheless, this correction scheme does not produce discontinuities. High-quality (CCSD(T) and DFT-SAPT) proton transfer potential energy curves are scarce, rendering such a parametrization difficult at this time. We plan to explore this avenue in the future.

2.3. Parameter Optimization. To improve the AM1 method for dispersion and hydrogen-bonding interactions, two empirical correction terms have been added as discussed above. These two correction terms involve a total of six parameters: two for the dispersion term (eq 1) and four for the H-bonding term (eq 4). These six parameters have been mathematically optimized to the RMSE of the interaction energies of 66 complexes (the F66 training set, see Table S1, Supporting Information). All of the interaction energies in the training set are CCSD(T) or SAPT quality. The training set consists of complexes not only at their minimum energy structures but also at greater and lesser separation than the potential minimum. Inclusion of these nonequilibrium structures is intended to increase the reliability of geometry optimization with AM1-FS1.

Our F66 training set includes the complexes in the S22 database,²⁰ which has been used by others for similar parametrization purposes.^{19,22,32} We have also included the four additional H-bonded complexes³³ that were later introduced to the S22 database, now termed the S26 database.

The additional interaction energies are also CCSD(T) quality. In our F66 training set, the water dimer, T-shaped benzene dimer, and both uracil dimer structures from the S22 database have been replaced by five points on their respective interaction potential energy curves. Five-point sampling of the potential energy curves has also been added for the nitromethane dimer,³⁴ parallel²³ and M1³⁵ benzene dimer, and three different benzene–acetylene³⁶ conformations. For a detailed list of complexes in the training set, refer to Table S1 in the Supporting Information. It would be desirable to have more potential energy curves in the training set, but there is limited high quality data available. For training set purposes, we have restricted ourselves to using only CCSD(T) or SAPT results, and only at or near the complete basis set limit.

Upon optimization of the parameters, the damping coefficient (d) in eq 2 optimized to infinity. This is because the AM1 method, as well as other semiempirical methods, inaccurately models repulsive interactions at close range for dispersion bound complexes. This can be observed by comparing DFT and semiempirical (AM1, PM3, RM1, and PM6) potential energy curves for the parallel benzene dimer, as shown in Figure 1. The figure clearly shows that at close separation the semiempirical methods (AM1, PM3, RM1, and PM6) differ significantly from the DFT (BLYP/6-311G(d,p)) results, severely underestimating the repulsion at close separations. The inaccurate repulsive inner wall of the potential is a consequence of the minimal basis set and parametrization of the SE methods.²⁵ This inaccuracy is the origin of d optimizing to infinity. As d becomes larger, the dispersion correction is turned off more rapidly; however, the function cannot become positive as needed to correct for underestimation of the repulsion at short range by the SE method. This problem could potentially be improved if a 6-12 LJ potential was used and only intercomponent atom pairs were considered; however, this introduces the requirement of atom connectivity information. It would also introduce a discontinuity in the potential and/or its derivative during bond breaking and formation processes (not to mention that intracomponent dispersion interactions would be neglected completely, thereby rendering the method ineffective for modeling conformational preference in macromolecules). We have therefore chosen to set d equal to 1000 and fully optimize the other five parameters. The damping coefficient was chosen to be 1000, because this is at the computational limit for evaluating the derivative of eq 1 within double precision. (Derivative information is needed for structural optimizations.) The other five parameters optimized to the following values: $S_6 = 1.1059$, $\alpha_1 = 0.4882$, $\alpha_2 = 0.6211$, $\alpha_3 = 0.3344$, and $\alpha_4 = 1.5451$.

2.4. Why Begin with AM1? AM1 has long been accepted as one of the most robust semiempirical methods. This method has been used many times with success for modeling large systems, but this is not the only reason for choosing AM1. We applied the same correction scheme described above to the RM1 method, which is a reparameterized version of AM1. The “corrected” RM1 method was actually less successful, on the basis of the RMSE for the F66 training set. Upon further investigation, we found that the RM1

method (as well as the PM3 and PM6 methods) performs worse than the AM1 method for the benzene dimer when compared to the DFT results that neglect dispersion interactions, as discussed above. Thus, if the same dispersion correction is applied to these mentioned SE methods, the AM1 method will produce the best result, even though other uncorrected SE methods produce potential energy curves closer to the CCSD(T) results. This is because the AM1 method has the strongest repulsive wall, therefore, producing a potential energy curve closest to the DFT result (see Figure 1). The functional form of the dispersion correction (eq 1) does not allow the term to become positive as is needed in some cases. This can be easily seen for the PM6 results in Figure 1. At close range, the CCSD(T) results are more repulsive than the PM6 results; thus to make the PM6 curve identical to the CCSD(T) curve, a repulsive correction would be needed. This problem is less severe for the AM1 method, rendering it more suitable for modeling dispersion interactions at close range. Note that these findings again might lead one to believe that using a function like the LJ potential would be beneficial, since a repulsive term is included. In fact, the LJ potential was among our many attempts to improve the AM1 method, but without success. This was due to the fact that we were/unwilling to add the burden of requiring atom connectivity information. We believe such a burden outweighs the potential added benefit.

3. Validation Studies

3.1. Single-Point Energies. In Table 1, the single-point interaction energies for the structures in the S22 database²⁰ are compared for various corrected and uncorrected SE methods. First, note that the interaction energies for the uncorrected AM1 and PM3 methods deviate significantly from the CCSD(T) reference values. The root-mean-square error (RMSE) for the AM1 and PM3 methods are 8.47 and 7.73 kcal/mol, respectively. Not only are these errors very large, but in many cases the sign of the interaction is predicted incorrectly. That is, the interactions are predicted to be repulsive not attractive. The addition of an empirical correction term(s) can drastically improve these methods. Our AM1-FS1 method reduces the RMSE to 1.18 kcal/mol, with the correct sign being predicted in all cases.

The results from McNamara and Hillier's¹⁹ AM1-D and PM3-D methods and the PM6-DH method of Řezáč et al.²¹ are also reported in Table 1. AM1-FS1 shows a slight improvement over the AM1-D method in two of the three subcategories and overall has a lower RMSE. AM1-FS1 achieves comparable accuracy to the PM3-D method for intermolecular interaction energies; the RMSEs are both 1.18 kcal/mol, with AM1-FS1 achieving a slightly lower MUE. While the overall improvement achieved by AM1-FS1 in the accuracy with which intermolecular binding energies are predicted is minor, we note that this has been achieved with significantly less parametrization and *no* modification of the original AM1 parameters.

The recently published PM6-DH method²¹ slightly outperforms AM1-FS1 on the basis of the single-point energies for the S22 database. Looking at the hydrogen bonded

complexes, the RMSEs are 1.07 and 1.37 kcal/mol for the PM6-DH and AM1-FS1 methods, respectively. Given that PM6-DH requires *different* parameters for each type of hydrogen bond, the 0.3 kcal/mol improvement in RMSE shown by PM6-DH is not especially significant. The group has identified eight H-bonding scenarios resulting in a total of 24 parameters for their H-bond correction term (three parameters for each H-bonding type). AM1-FS1 only uses four parameters; AM1-FS1 also does not introduce the requirement of knowing atom conductivity. The PM6-DH method does show significant improvement for many dispersion bonded cases, but it performs poorly for modeling the potential energy surface of the benzene dimer. This is discussed below and shown graphically in section 3.4. It should be noted that Řezáč et al.²¹ only used complexes 8–22 of the S22 database for determining the dispersion parameters for PM6-DH. Thus, it is not unexpected that good agreement was achieved for the eight dispersion bound complexes.

3.2. F66 Results. Many of the other empirically corrected SE methods discussed have been parametrized to the S22 database, thus they should achieve accurate results for those complexes. AM1-FS1 has been parametrized to a larger training set consisting of 66 complexes. Parameterizing to this larger training set has led to an increase in RMSE for the S22 database. This is not unexpected and, in our opinion, is a worthwhile sacrifice that should make AM1-FS1 more versatile. (In fact, we tried optimizing AM1-FS1 solely to the S22 database and achieved near DFT-D level accuracy, but when the method was subsequently tested on the F66 training set, a larger RMSE resulted.) AM1-FS1 is parametrized to the F66 training set and achieves a sub-kilocalorie RMSE (0.99 kcal/mol) and MUE (0.69 kcal/mol). The individual results are reported in Table S1 in the Supporting Information. Both AM1-D and PM3-D were parametrized solely to the S22 database, so high accuracy is not surprising when the S22 database is used as the “test set”. We have performed calculations on the 66 complexes of the F66 set using McNamara and Hillier's AM1-D method for comparison. This provides for a much more comprehensive test of the method than the S22 database because it contains a wider variety of structures *and* nonequilibrium structures. AM1-D produces a RMSE and MUE of 1.49 and 1.02 kcal/mol respectively, approximately 50% less accurate than AM1-FS1. Upon close inspection of Table S1, it can be seen that AM1-FS1 significantly outperforms AM1-D on the repulsive wall, an issue we will look at more closely in section 3.4.

3.3. Optimized Energies and Structures. This section considers the effect of geometry optimization on interaction energy and structural distortion for systems in the S22 database. The ability of an empirically corrected SE method to perform accurately in this role is crucial because one of the principal uses of SE methodology is structural optimization of systems that are too large for optimization with first-principles methods. The ability of a method to reproduce interaction energies at *reference* geometries is not very useful, because if we know the CCSD(T) geometry, and therefore its energy, there is little value in knowing the SE energy for that structure. In Table 2, the interaction energies for the

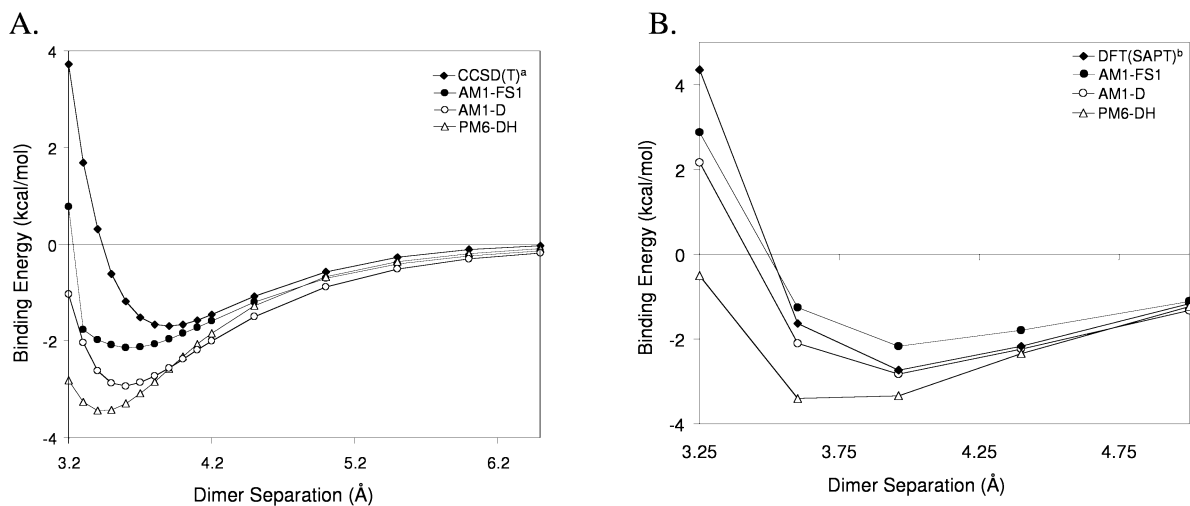


Figure 3. Parallel and M1 benzene dimer potential energy curves determined with various computational methods. Superscript a refers to ref 23 and b to ref 35.

geometry optimized complexes are reported for a variety of corrected and uncorrected SE methods. Again, the uncorrected AM1 and PM3 methods perform poorly. Upon applying our correction scheme to the AM1 method, the RMSE is lowered from 6.25 to 1.80 kcal/mol. AM1-FS1 is also an improvement over AM1-D; AM1-FS1 outperforms AM1-D in all subcategories by about 1 kcal/mol. This increase in performance for optimization, compared to the AM1-D method, presumably results from our use of a substantially larger training set that includes nonequilibrium structures.

The performance of McNamara and Hillier's PM3-D method is comparable to our AM1-FS1 method. Depending on the statistical metric selected, either may be said to outperform the other for predicting interaction energies upon geometry optimization of the structure in the S22 database. AM1-FS1 does perform better in two of the three categories, the dispersion and mixed bounded complexes. The PM6-DH method outperforms all the other methods. However, structural distortion should also be considered, but unfortunately, data for such a comparison are not available. Again, this aspect of a SE method is especially important since such a method will likely be used for optimization purposes.

To gauge the degree of structural distortion upon geometry optimization, select interaction distances are compared and are shown in Table 3. The interaction distances are defined as the center-of-mass separation and/or atom-atom distance(s) between the two monomers depending on the system (see Figure S1 of ref 19 for the specific interaction distances). Comparing the different empirically corrected SE methods, we find that AM1-FS1 outperforms AM1-D in every category. Our method is generally comparable to PM3-D on the basis of interaction distance. AM1-FS1 performs better in two of the three categories. This time, AM1-FS1 outperforms PM3-D for the H-bonded complexes on the basis of interaction distances, but not for the dispersion bound complexes. Based solely on the total RMSE for the S22 database would be difficult to choose which method, AM1-FS1 or PM3-D, is better; however, AM1-FS1 does not require reoptimization of the AM1 parameters thereby

preserving the predictive power of AM1 for calculation of heats of formation, discussed below. As noted above, interaction distances and/or structural geometries were not made available for the PM6-DH method preventing structural comparisons upon optimization of the S22 complexes.

To further test the ability of AM1-FS1 to model H-bonding complexes, 16 additional hydrogen bonded DNA base pairs have been considered. The 16 additional complexes were chosen from ref 20 since these are the only complexes from the H-bonding subsection that have CCSD(T) quality binding energies. The geometries of these complexes, however, are from MP2 optimizations or experimental data. Therefore, these structures do not correspond to the CCSD(T) potential minimum; this is also the case for most of the S22 database structures. We have computed the binding energies for these complexes on the basis of the reference geometries and also AM1-FS1 optimized geometries. The RMSEs for the binding energies are 1.78 and 2.18 kcal/mol, respectively. This error is consistent with the error associated with the hydrogen bonding complexes in the S22 database, which were used for parametrization. The 16 complexes as well as the reference CCSD(T), single point, and optimized AM1-FS1 binding energies are reported in Table S2 of the Supporting Information.

3.4. Potential Energy Curves. The value of a computational method is significantly enhanced if it is able to accurately describe the potential energy surface apart from the minimum. A given method could accurately predict the interaction energy at a specific molecular geometry yet yield a very inaccurate picture of the remainder of the potential energy surface. (See ref 25 for a detailed discussion.) In this section, potential energy curves will be compared for various empirically corrected SE methods.

In Figure 3, potential energy curves for two different benzene dimer conformations are shown. Figure 3A shows the interaction energy for the parallel dimer as a function of monomer separation. The parallel dimer is not the lowest energy conformation, but it is important to be able to model a variety of geometries correctly for the correct description of π - π interactions involved in large systems, and the

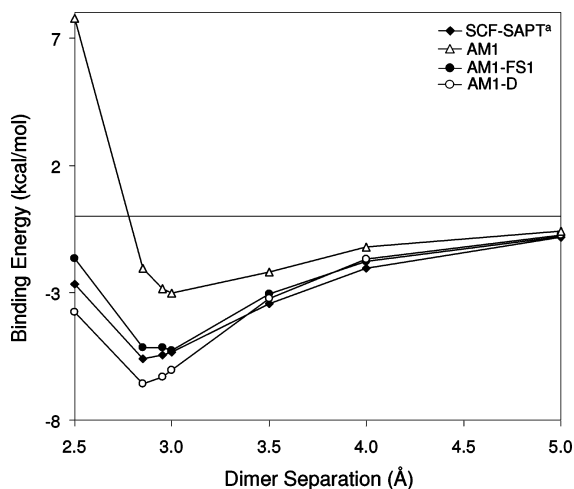


Figure 4. Water dimer potential energy curves as a function of O---O separation, as determined with various computational methods. Superscript a refers to ref 37.

parallel dimer represents a widely used test case, probably owing to the simplicity of its construction. The M1 benzene dimer, according to ref 35, is the lowest energy structure known. Compared to the CCSD(T) and DFT(SAPT) reference values, among the empirically corrected SE methods PM6-DH performs the worst for these systems. The PM6-DH method seems to overbind π - π interactions. This is due to the fact that the PM6 method performs poorly for dispersion bound complexes compared to DFT-BLYP results as discussed earlier and shown in Figure 1 (further discussion is presented in ref 25). McNamara and Hillier's AM1-D method performs very well for the M1 dimer, however, not so well for the parallel dimer. On average, AM1-FS1 performs the best for these two systems. AM1-FS1 has a very steep potential wall at close separation (Figure 3A); this is an artifact of using a large damping parameter (d) in the dispersion correction term (eq 2). Again, the large term is required because of the inability of the AM1 method to properly capture short-range repulsive interactions.

In Figures 4 and 5, potential energy curves for the water dimer and the nitromethane dimer are shown, respectively. The water dimer is a classic hydrogen bonding system. The potential energy curves in Figure 4 are shown as a function of O---O separation. The figure shows that AM1-FS1 dramatically improves upon the AM1 method and outperforms McNamara and Hillier's AM1-D method. The correlation to SCF-SAPT results³⁷ again shows that the hydrogen bonding correction term (eq 4) is a worthwhile addition to the AM1 method. The AM1-D method also performs relatively well for the water dimer. This means that the changes they have made to the AM1 parameters improve the results for this particular system; however, the same is not observed if we consider the nitromethane dimer. In Figure 5, we see that AM1-D performs poorly for the nitromethane dimer. The SCF-SAPT curve³⁴ is for the so-called "double hydrogen bond" configuration; however, nitromethane is not a classical H-bonding system. It lacks a hydrogen atom attached to a highly electronegative atom (N, O, or F); nevertheless, this system is said to form weak H bonds.³⁴ As shown (Figure 5), the AM1 method performs relatively

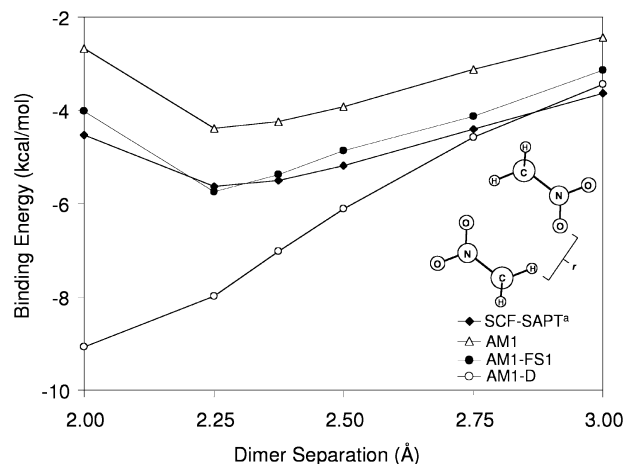


Figure 5. Nitromethane dimer potential energy curve in the "double hydrogen bond" configuration, as determined with various computational methods. Superscript a refers to ref 34.

well for this system, whereas McNamara and Hillier's modification of the AM1 parameters has caused the AM1-D method to inadequately model this system. AM1-FS1, on the other hand, does not consider this a H-bonding case. Therefore, the H-bonding correction term is not turned on for this system. Consequently, AM1-FS1 performs well for this system by applying only the dispersion correction. This potential energy curve demonstrates that the AM1 parameters should not be changed in all cases. It should be noted that the AM1-D training set does not contain this system, whereas the training set for AM1-FS1 does. (We have not compared the PM6-DH method of Řezáč et al.²¹ in H-bonding cases because we do not have code for their elaborate H-bonding correction scheme.)

3.5. Heat of Formation. As mentioned earlier, modifying the original semiempirical parameters can have deleterious effects, especially for thermodynamic properties. For example, the experimental heat of formation of benzene is 19.8 kcal/mol³⁸ and is predicted to be 22.0 kcal/mol by the AM1 method.¹⁵ The AM1-D method, however, predicts a value of -12.9 kcal/mol. (PM3-D performs even more poorly, yielding -21.8 kcal/mol.) Reparameterization has rendered AM1-D (and PM3-D) unreliable for predicting thermodynamic properties. On the other hand, AM1-FS1 does not change any of the original AM1 parameters and predicts the heat of formation of benzene to be 20.0 kcal/mol, in good agreement with experimental results and, serendipitously, even a slight improvement over AM1. The AM1-FS1 empirical correction is designed to have little effect on quantities that are already predicted relatively well by the AM1 method. Table 4 collects results for calculations of heat of formation on 53 test molecules. Note that the RMSE in predictions of heat of formation with AM1-FS1 is comparable to that of the original AM1 method, but AM1-D is 24 times (2400%) less accurate. Reparameterization of the original PM3 method in the development of PM3-D has also seriously degraded its predictive power for heats of formation (see Table 4). This clearly shows the negative consequences of changing the original semiempirical parameters without

Table 4. Heat of Formation (kcal/mol)^a

molecule	heat of formation (kcal/mol)				
	expt	AM1	AM1-D	PM3-D	AM1-FS1
methane	-17.8	-8.8	-78.4	-6.6	-8.8
ethane	-20.0	-17.4	-114.1	-12.2	-18.3
ethylene	12.5	16.5	-39.4	-11.0	16.1
acetylene	54.5	54.8	38.6	-9.8	54.8
propane	-25.0	-24.3	-148.4	-17.7	-27.1
propene	4.8	6.6	-75.6	-16.5	5.3
propyne	44.2	43.4	3.4	-15.3	43.0
allene	45.5	46.1	6.6	-15.3	45.7
<i>n</i> -butane	-30.0	-31.1	-182.7	-23.2	-36.1
isobutane	-32.0	-29.4	-181.1	-23.2	-35.3
but-1-ene	-0.1	0.4	-109.2	-22.0	-2.6
<i>trans</i> -2-butene	-2.8	-3.3	-111.8	-22.0	-5.6
<i>cis</i> -2-butene	-1.7	-2.2	-110.9	-22.0	-4.8
isobutene	-4.0	-1.2	-109.8	-22.0	-4.3
1,2-butadiene	38.8	37.1	-28.5	-20.8	35.9
<i>trans</i> -1,3-butadiene	26.3	29.9	-38.2	-20.8	28.3
1-butyne	39.5	37.5	-29.8	-20.8	35.6
2-butyne	34.8	32.0	-31.8	-20.8	31.1
vinylacetylene	72.8	67.9	42.1	-19.6	66.9
diacetylene	113.0	106.1	122.4	-18.4	105.8
<i>n</i> -pentane	-35.1	-37.9	-216.9	-28.7	-45.1
neopentane	-40.2	-32.8	-212.3	-28.7	-42.8
benzene	19.8	22.0	-12.9	-29.6	20.0
toluene	12.0	14.5	-46.6	-35.1	10.7
ammonia	-11.0	-7.3	-154.0	-7.7	-7.3
methylamine	-5.5	-7.4	-165.2	-13.3	-8.2
dimethylamine	-4.4	-5.6	-175.1	-18.8	-7.5
trimethylamine	-5.7	-1.7	-183.3	-24.3	-5.1
ethylamine	-11.3	-15.1	-200.4	-18.8	-17.5
<i>n</i> -propylamine	-16.8	-22.1	-234.7	-24.3	-26.5
isopropylamine	-20.0	-19.2	-231.9	-24.3	-23.9
<i>tert</i> -butylamine	-28.9	-21.2	-261.6	-29.8	-29.3
pyrrole	25.9	39.9	-56.0	-26.3	38.5
pyridine	34.6	32.1	-29.5	-30.7	30.6
pyridazine	66.5	55.3	-33.6	-31.8	54.2
water	-57.8	-59.2	-200.8	-12.0	-59.2
methanol	-48.2	-57.0	-193.7	-17.5	-57.5
ethanol	-56.2	-62.7	-225.8	-23.0	-64.4
1-propanol	-61.0	-70.6	-261.9	-28.5	-74.7
2-propanol	-65.2	-67.7	-258.2	-28.5	-71.8
<i>t</i> -butyl_alcohol	-74.7	-71.6	-288.7	-34.0	-78.6
dimethyl_ether	-44.0	-53.2	-185.7	-23.0	-53.8
diethyl_ether	-60.3	-64.4	-249.6	-34.0	-67.8
oxirane	-12.6	-8.9	-95.3	-21.8	-9.2
furan	-8.3	3.0	-52.7	-30.5	2.1
phenol	-23.0	-22.2	-120.0	-40.4	-24.9
anisole	-16.2	-15.8	-110.1	-45.9	-19.8
benzaldehyde	-8.8	-8.9	-58.5	-44.8	-12.1
formic_acid	-90.5	-97.4	-222.7	-27.2	-97.4
acetic_acid	-103.4	-103.0	-252.9	-32.7	-103.7
propionic_acid	-108.4	-108.0	-285.5	-38.2	-111.1
oxalic_acid	-173.0	-172.4	-370.9	-53.2	-172.8
benzoic_acid	-70.3	-68.0	-181.4	-55.6	-71.0
RMSE		4.9	132.5	46.7	5.6
MUE		3.6	118.4	36.1	4.3

Binding Energies (kcal/mol) of Carbon Nanostructure Complexes^c

	AM1-FS1	AM1-D	M06-2X/6-31+G(d,p)// M06-L/MIDI
HMB@6CPPA	-16.6	-17.6	-14.7
C60@6CPPA	-26.9	-30.1	-28.0
C70@6CPPA	-36.3	-41.0	-31.1
3,3@6CPPA	-17.7	-22.1	-5.4
4,4@6CPPA	-32.7	-42.0	-24.0
5,5@6CPPA	-43.2	-46.4	-43.3
C60@BuckyCatcher ^b	-29.3	-36.8	-26.4
C60@Coroanulene ^b	-13.4	-16.7	-12.4

^a The experimental and AM1 results were obtained from ref 15. Note: the AM1-D and PM3-D results were obtained by coding the method outlined in ref 19; however, slight disagreements in the binding energies were observed with PM3-D for compounds containing oxygen, suggesting a misprint in the published PM3-D oxygen parameters. ^b Results were obtained from ref 45. ^c The M06 results were obtained from ref 44.

the consideration of such quantities during optimization of the parameters.

While the original AM1 parameters have not been altered in AM1-FS1, the FS1 correction terms do influence the predicted heat of formation. This occurs because the heat of formation is in part determined from the total energy of the complex, which now contains the empirical correction energy, but also in part from the energies of the isolated atoms. The isolated atom energies do not include any empirical correction energy since; by design, the correction terms are not implemented for a single atom. Therefore, the difference between the heat of formation as computed with AM1 and AM1-FS1 will generally become larger as the correction term(s) contribution increases. This will also be the case for DFT-D methods when the total energy is used in the determination of the heat of formation. The influence of the FS1 correction on the predicted heat of formation can have both undesirable and desirable consequences. For example, in Table 4, it can be seen that as the number of methylene units in the aliphatic hydrocarbons increases (methane → ethane → propane → etc.), the error in the predicted heat of formation increases. Fortunately, since the original AM1 parameters have not been altered, the AM1 heat of formation can be easily obtained by subtracting out the empirical correction energy from the AM1-FS1 heat of formation. This approach of subtracting the correction energy will also be effective for DFT-D methods when the total energy is used in the determination of the heat of formation. On the other hand, the correction to the total energy sometimes has a beneficial impact on the predicted heat of formation. For example, the experimental heat of formation of the benzene dimer is 30.4 kcal/mol³⁹ and is predicted to be 37.9 and 44.1 kcal/mol on the basis of structures optimized with the AM1-FS1 and AM1 methods, respectively. Note that these structures are significantly different upon optimization because the AM1 method does not consider dispersion interactions. If we determine the heat of formation of the AM1-FS1 optimized geometry with the AM1 method, the error is even larger; the heat of formation is predicted to be 46.7 kcal/mol. Since the original AM1 parameters were optimized to give reliable predictions of the heat of formation at AM1-optimized geometries, it seems reasonable to conclude that, in general, if the heat of formation of some large multicomponent carbon structure is desired, the AM1-FS1 method will likely produce a more accurate result at the AM1-FS1 geometry. Certainly, the AM1-FS1 structure will be more accurate since the dominant intercomponent interaction will be incorporated. This situation is much less complicated when a heat of reaction is of interest since the correction term(s) are applied to both the reactants and products.

4. Application to Macromolecular Complexes

The ultimate goal of AM1-FS1 is to be able to efficiently and accurately model large weakly bound systems, such as complexes of carbon nanostructures and molecular devices. Such large systems are currently out of reach for CCSD(T), and extremely computationally demanding for DFT methods. Furthermore, most DFT functionals are incapable of model-

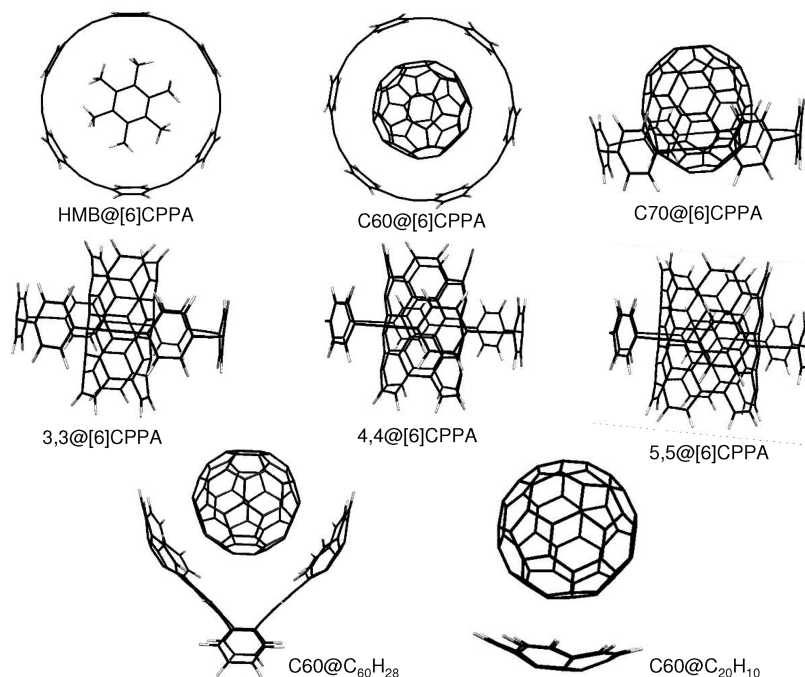


Figure 6. AM1-FS1 geometry optimized carbon nanostructure complexes.

ing carbon nanostructure complexes due to the fact that these systems are governed by van der Waals interactions. Presumably, the current most accurate methods capable of modeling such systems are DFT-D methods and the M05,⁴⁰ M06,⁴¹ and M08⁴² family of functionals developed by Zhao and Truhlar. Performing geometry optimizations with DFT-based methods on systems larger than 100 atoms is currently extremely computationally expensive, but such optimizations can be routinely performed with semiempirical-based techniques even on “PC”-class computers. This great computational efficiency of semiempirical methods provides the central motivation for the present work.

4.1. Carbon Nanostructures. To test the performance of AM1-FS1 on complexes of carbon nanostructures, we have performed geometry optimizations and determined the binding energy of several inclusion complexes. The hosts considered are corannulene ($C_{20}H_{10}$), a double-concave hydrocarbon buckycatcher⁴³ ($C_{60}H_{28}$), and cyclic[6]paraphenylacetylene (6CPPA). The AM1-FS1 optimized structures are shown in Figure 6. (It is important to note that these complexes would be predicted to be unbound if the standard AM1 and most DFT methods were used.) To date, the best binding energy values for these complexes are from DFT calculations reported by Zhao and Truhlar,^{44,45} using the M06-2X functional. The binding energies along with the AM1-FS1 and AM1-D results are reported in Table 4. The AM1-FS1 results are very comparable to the M06 values; however, AM1-FS1 overestimates the binding energy for 3,3@[6]CPPA and 4,4@[6]CPPA in comparison to DFT-M06-2X. This may be a result of the M06 functional underestimating dispersion interactions at long range. This hypothesis is supported by the potential energy curve for the parallel benzene dimer shown in Figure 7, which clearly shows that the M06 functional fails to accurately model dispersion interactions in the long-range regime, where the predicted interaction energy even becomes slightly positive.

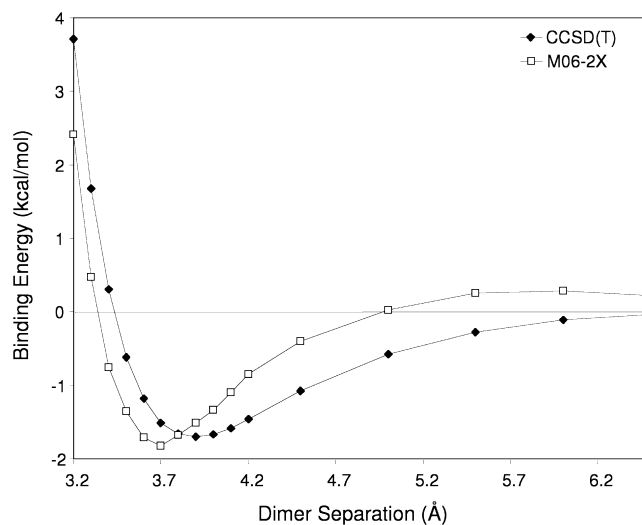


Figure 7. Parallel benzene dimer potential energy curve calculated with the M06-2X functional, using the 6-311G(d,p) basis. CCSD(T) results were obtained from ref 23.

This behavior might be easily overlooked since upon optimization of the parallel benzene dimer, a reasonable energy and structure will be produced. To show that this is the case for 3,3@[6]CPPA, the binding energy was determined using the BLYP-D functional. The resulting binding energy of 18.9 kcal/mol is in very good agreement with the AM1-FS1 result. The M06-2X functional underestimates binding for 3,3@[6]CPPA because the nearest intermolecular interaction is 4.5 Å, a distance at which M06 underestimates the interaction energy as exhibited by the benzene dimer potential energy curve.

The AM1-FS1 method significantly outperforms AM1-D in every case on the basis of the current benchmark M06 values. AM1-FS1 achieves this correlation with only two added parameters (due to the nature of the systems only the

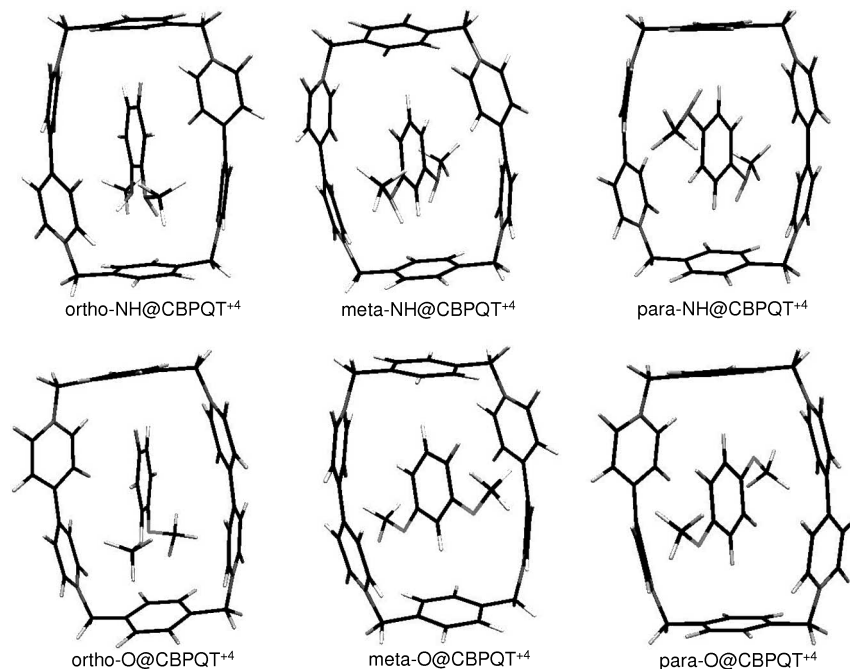


Figure 8. AM1-FS1 geometry optimized pseudorotaxane complexes.

Table 5. Binding Energies (kcal/mol) of Pseudorotaxane Complexes^a

	Binding Energy (kcal/mol)		
	AM1-FS1	M06-2X/6-311G(d,p)//M06-L/MIDI	LMP2/6-311+G(d,p)//BHandHLYP/6-31G(d)
ortho-O@CBPQT ⁺⁴	-32.1	-34.7	-21.0
meta-O@CBPQT ⁺⁴	-31.7	-35.0	-16.1
para-O@CBPQT ⁺⁴	-33.4	-34.7	-21.3
ortho-NH@CBPQT ⁺⁴	-37.7	-41.7	-22.3
meta-NH@CBPQT ⁺⁴	-38.2	-36.9	-22.5
para-NH@CBPQT ⁺⁴	-38.5	-40.1	-23.9

^a The LMP2 results were obtained from ref 46.

dispersion correction term is “turned on” during the AM1-FS1 calculations), whereas AM1-D utilizes 10 parameters. We credit the success of AM1-FS1 to parametrizing to a larger training set containing nonequilibrium complexes.

4.2. Pseudorotaxanes. We also tested the performance of AM1-FS1 on six different pseudorotaxanes, since these types of complexes are of central interest to our research group. All of the systems considered incorporate cyclobis-(paraquat-p-phenylene) (CBPQT⁺⁴), a tetracationic ring structure. Six inclusion complexes with this ring have been formed with dimethoxybenzene and benzenedimethanamine in the *ortho*, *meta*, and *para* conformations. (AM1-FS1 optimized structures are shown in Figure 8.) We have performed geometry optimizations and determined the binding energies of these complexes and compared them to previously reported LMP2/6-311+G(d,p)//BHandHLYP/6-31G(d) results.⁴⁶ We also computed these binding energies at the M06-2X/6-311G(d,p)//M06-L/MIDI level of theory for additional comparisons. (All results are reported in Table 5.) On the basis of the results, the LMP2/6-311+G(d,p)//BHandHLYP/6-31G(d) results appear to underestimate the binding energy. This is likely a result of the geometry produced by the BHandHLYP functional and not the LMP2

method. This conclusion is based on the binding energies determined at the M06-2X/6-311G(d,p)//M06-L/MIDI level. The differences in binding energies between the isomers are sufficiently small that they may be taken to be insignificant given the level of theory and the large conformational space associated with these complexes. On the basis of these results, we believe AM1-FS1 is a valuable tool for modeling this class of macromolecular complexes.

5. Conclusions

AM1-FS1 is a new empirically corrected semiempirical method suitable for performing geometry optimizations on macromolecular complexes. AM1-FS1 displays considerable improvement over the traditional AM1 method for nonbonding interactions, yet it retains the computational efficiency and predictive power for thermochemical quantities of the original AM1 Hamiltonian. Validation testing shows that the method reduces the RMSE for the popular S22 database from 8.47 to 1.18 kcal/mol. More impressively, this new method has achieved kilocalorie accuracy on a training set of 66 complexes. This was accomplished with just six empirical parameters (two for dispersion and four for hydrogen-bonding) and *no* reparameterization of AM1 (which we show here has led to serious consequences in existing empirically corrected SE methods). This is a dramatic reduction in the total number of adjustable parameters compared to other previously published empirically corrected SE methods. Validation testing shows that, while the existing PM6-DH method does outperform AM1-FS1 on the basis of the S22 database, PM6-DH is shown to be inaccurate for reproducing potential energy curves for the benzene dimer, a classic test case used for predicting the likely accuracy of a method for modeling complexes of carbon nanostructures. Moreover, unlike PM6-DH, AM1-FS1 does *not* require knowledge of atom connectivity. On the basis of the examples reported,

on average, AM1-FS1 is also the most reliable empirically corrected SE method for reproducing the potential energy curve away from the global minimum. We credit this success to using a training set that contains nonequilibrium complexes.

This new AM1-FS1 method has been shown to yield results comparable in accuracy to the best available calculations on complexes of carbon nanostructures and carbohydrate pseudorotaxanes. We believe AM1-FS1 is a useful computational tool for obtaining reliable results for such systems at limited computational expense. It should prove to be a valuable asset for routine modeling of macromolecular complexes that are currently at (or beyond) the limit of DFT based techniques, or out of reach of higher levels of theory.

Acknowledgment. This work was supported by National Science Foundation grant CHE0449595 and E. I. du Pont de Nemours & Co., Inc.

Supporting Information Available: The complete F66 training set is reported along with the interaction energies and appropriate references. The interaction energies for the F66 complexes are reported for AM1-FS1 and McNamara and Hillier's¹⁹ AM1-D method. This information is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Vallee, R.; Damman, P.; Dosiere, M.; Toussaere, E.; Zyss, J. Nonlinear Optical Properties and Crystalline Orientation of 2-Methyl-4-nitroaniline Layers Grown on Nanostructured Poly(tetrafluoroethylene) Substrates. *J. Am. Chem. Soc.* **2000**, *122*, 6701–6709.
- Thalladi, V. R.; Brasselet, S.; Weiss, H.-C.; Blaser, D.; Katz, A. K.; Carrell, H. L.; Boese, R.; Zyss, J.; Nangia, A.; Desiraju, G. R. Crystal Engineering of Some 2,4,6-Triaryloxy-1,3,5-triazines: Octupolar Nonlinear Materials. *J. Am. Chem. Soc.* **1998**, *120*, 2563–2577.
- Hunter, C. A.; Sanders, J. K. M. The nature of pi-pi interactions. *J. Am. Chem. Soc.* **1990**, *112*, 5525–5534.
- Xiao, Y.; Chen, C.; He, Y. Folding Mechanism of Beta-Hairpin Trpzip2: Heterogeneity, Transition State and Folding Pathways. *Int. J. Mol. Sci.* **2009**, *10*, 2838–2848.
- Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. *Proteins: Str. Funct. Genet.* **2002**, *47*, 409–443.
- Jonikas, M. C.; Collins, S. R.; Denic, V.; Oh, E.; Quan, E. M.; Schmid, V.; Weibezahn, J.; Schwappach, B.; Walter, P.; Weissman, J. S.; Schuldiner, M. Comprehensive Characterization of Genes Required for Protein Folding in the Endoplasmic Reticulum. *Science* **2009**, *323*, 1693–1697.
- Dobson, C. M. Protein folding and misfolding. *Nature* **2003**, *426*, 884–890.
- Griffiths-Jones, S. R.; Searle, M. S. Structure, Folding, and Energetics of Cooperative Interactions between the beta-Strands of a *de Novo* Designed Three-Stranded Antiparallel beta-Sheet Peptide. *J. Am. Chem. Soc.* **2000**, *122*, 8350–8356.
- Burley, S. K.; Petsko, G. A. Aromatic-Aromatic Interaction: A Mechanism of Protein Structure Stabilization. *Science* **1985**, *229*, 23–28.
- Guckian, K. M.; Krugh, T. R.; Kool, E. T. Solution Structure of a Nonpolar, Non-Hydrogen-Bonded Base Pair Surrogate in DNA. *J. Am. Chem. Soc.* **2000**, *122*, 6841–6847.
- Foster, M. E.; Sohlberg, K. Theoretical Study of Binding Site Preference in ²Rotaxanes. *J. Chem. Theory Comput.* **2007**, *3*, 2221–2233.
- Zheng, X.; Sohlberg, K. Modeling bistability and switching in a ²catenane. *Phys. Chem. Chem. Phys.* **2004**, *6*, 809–815.
- Nepogodiev, S. A.; Stoddart, J. F. Cyclodextrin-Based Catenanes and Rotaxanes. *Chem. Rev.* **1998**, *98*, 1959–1976.
- Zheng, X.; Sohlberg, K. Modeling of a Rotaxane-based Molecular Device. *J. Phys. Chem.* **2003**, *107*, 1207–1215.
- Dewar, M. J. S.; Zorbisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.
- Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RM1: A Reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27*, 1101–1111.
- Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.
- McNamara, J. P.; Hillier, I. H. Semi-empirical molecular orbital methods including dispersion corrections for the accurate prediction of the full range of intermolecular interactions in biomolecules. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362–2370.
- Jurečka, P.; Spöner, J.; Cerný, J.; Hobza, P. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. Semiempirical Quantum Chemical PM6Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.
- Jurečka, P.; Cerný, J.; Hobza, P.; Salahub, D. Density Functional Theory Augmented with an Empirical Dispersion Term. Interaction Energies and Geometries of 80 Noncovalent Complexes Compared with Ab Initio Quantum Mechanics Calculations. *J. Comput. Chem.* **2007**, *28*, 555–569.
- Sinnokrot, M. O.; Sherrill, C. D. Highly Accurate Coupled Cluster Potential Energy Curves for the Benzene Dimer: Sandwich, T-Shaped, and Parallel-Displaced Configurations. *J. Phys. Chem. A* **2004**, *108*, 10200–10207.
- Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- Foster, M. E.; Sohlberg, K. Empirically corrected DFT and semi-empirical methods for non-bonding interactions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 307–322.
- Grimme, S. Accurate Description of van der Waals Complexes by Density Functional Theory Including Empirical Corrections. *J. Comput. Chem.* **2004**, *25*, 1463–1473.
- Wu, Q.; Yang, W. Empirical correction to density functional theory for van der Waals interactions. *J. Chem. Phys.* **2002**, *116*, 515–524.

- (28) Halgren, T. A. Representation of van der Waals (vdW) Interaction in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and vdW Parameters. *J. Am. Chem. Soc.* **1992**, *114*, 7827–7843.
- (29) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (30) Chirgwin, H.; Coulson, C. A. The Electronic Structure of Conjugated Systems. VI. *Proc. R. Soc. London, Ser. A* **1950**, *201*, 196–209.
- (31) Buckingham, A. D.; Bene, J. E. D.; McDowell, S. A. C. The hydrogen bond. *Chem. Phys. Lett.* **2008**, *463*, 1–10.
- (32) Schwabe, T.; Grimme, S. Double-hybrid density functionals with long-range dispersion corrections: higher accuracy and extended applicability. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.
- (33) Riley, K. E.; Hobza, P. Assessment of the MP2 Method, along with Several Basis Sets, for the Computation of Interaction Energies of Biologically Relevant Hydrogen Bonded and Dispersion Bound Complexes. *J. Phys. Chem. A* **2007**, *111*, 8257–8263.
- (34) Cole, S. J.; Szalewicz, K., III; Bartlett, R. J. Correlated calculation of the interaction in the nitromethane dimer. *J. Chem. Phys.* **1986**, *84*, 6833–6836.
- (35) Podszwa, R.; Bukowski, R.; Szalewicz, K. Potential Energy Surface for the Benzene Dimer and Perturbational Analysis of π - π Interactions. *J. Phys. Chem. A* **2006**, *110*, 10345–10354.
- (36) Tekin, A.; Jansen, G. How accurate is the density functional theory combined with symmetry-adapted perturbation theory approach for CH- π and π - π interactions? A comparison to supermolecular calculations for the acetylene-benzene dimer. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1680–1687.
- (37) Rybak, S.; Jeziorski, B.; Szalewicz, K. Many-body symmetry-adapted perturbation theory of intermolecular interactions. H₂O and HF dimers. *J. Chem. Phys.* **1991**, *95*, 6576–6601.
- (38) Pedley, J. B.; Rylance, J. *Sussex-N.P.L. Computer Analysed Thermochemical Data: Organic and Organometallic Compounds*; University of Sussex: East Sussex, U.K., 1977.
- (39) Grover, J. R. Dissociation Energies of the Benzene Dimer and Dimer Cation. *J. Phys. Chem.* **1987**, *91*, 3233–3237.
- (40) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. Design of Density Functionals by Combining the Method of Constraint Satisfaction with Parametrization for Thermochemistry, Thermochemical Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.
- (41) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (42) Zhao, Y.; Truhlar, D. G. Exploring the Limit of Accuracy of the Global Hybrid Meta Density Functional for Main-Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2008**, *4*, 1849–1868.
- (43) Sygula, A.; Fronczek, F. R.; Sygula, R.; Rabideau, P. W.; Olmstead, M. M. A Double Concave Hydrocarbon Bucky-catcher. *J. Am. Chem. Soc.* **2007**, *129*, 3842–3843.
- (44) Zhao, Y.; Truhlar, D. G. Size-selective supramolecular chemistry in a hydrocarbon nanoring. *J. Am. Chem. Soc.* **2007**, *129*, 8440–8442.
- (45) Zhao, Y.; Truhlar, D. G. Computational Characterization and modeling of buckyball tweezers: density functional study of concave-convex π - π interactions. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2813–2818.
- (46) Romero, C.; Fomina, L.; Fomone, S. How important is the dispersion interaction for cyclobis(paraquat-p-phenylene)-based molecular “shuttles”? A theoretical study. *Int. J. Quantum Chem.* **2005**, *102*, 200–208.
- (47) Pitonak, M.; Riley, K. E.; Neogrady, P.; Hobza, P. Highly Accurate CCSD(T) and DFT-SAPT Stabilization Energies of H-Bonded and Stacked Structures of the Uracil Dimer. *ChemPhysChem* **2008**, *9*, 1636–1644.

CT100177U

Molecular Dynamics in Physiological Solutions: Force Fields, Alkali Metal Ions, and Ionic Strength

Chao Zhang,[†] Simone Raugei,[‡] Bob Eisenberg,[§] and Paolo Carloni^{*,¶,||}

German Research School for Simulation Sciences, FZ-Juelich/RWTH Aachen University, Aachen, Germany, Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, Washington 99352, Rush University Medical Center, 1653 W. Congress Parkway, Chicago, Illinois 60612, and SISSA, CNR-INFN-DEMOCRITOS, and Italian Institute of Technology (IIT), SISSA Unit, Trieste, Italy

Received December 6, 2009

Abstract: The monovalent ions Na^+ and K^+ and Cl^- are present in any living organism. The fundamental thermodynamic properties of solutions containing such ions is given as the excess (electro-)chemical potential differences of single ions at finite ionic strength. This quantity is key for many biological processes, including ion permeation in membrane ion channels and DNA–protein interaction. It is given by a chemical contribution, related to the ion activity, and an electric contribution, related to the Galvani potential of the water/air interface. Here we investigate molecular dynamics based predictions of these quantities by using a variety of ion/water force fields commonly used in biological simulation, namely the AMBER (the newly developed), CHARMM, OPLS, Dang95 with TIP3P, and SPC/E water. Comparison with experiment is made with the corresponding values for salts, for which data are available. The calculations based on the newly developed AMBER force field with TIP3P water agrees well with experiment for both KCl and NaCl electrolytes in water solutions, as previously reported. The simulations based on the CHARMM-TIP3P and Dang95-SPC/E force fields agree well for the KCl and NaCl solutions, respectively. The other models are not as accurate. Single cations excess (electro-)chemical potential differences turn out to be similar for all the force fields considered here. In the case of KCl, the calculated electric contribution is consistent with higher level calculations. Instead, such agreement is not found with NaCl. Finally, we found that the calculated activities for single Cl^- ions turn out to depend clearly on the type of counterion used, with all the force fields investigated. The implications of these findings for biomolecular systems are discussed.

1. Introduction

Monovalent ions, such as Na^+ and K^+ and Cl^- , are essential to life. For example, the name of the channel protein that

conducts these ions across the membranes of cells is often given by its selectivity for single ions (e.g., sodium, potassium, and chloride channels). All living processes occur in the presence of the electrolyte solution with finite ionic strength: solutions outside cells are mostly Na^+ (about 0.14 molal or m)¹ and inside cells mostly K^+ and Cl^- (0.14 and 0.1 m, respectively).² Ions move through selective channels,³ where local ionic strength can be as large as 5 m^{4,5} and rearrange dramatically in the formation of protein–, DNA–, and RNA–protein complexes.^{6–8} Therefore, the thermodynamics of *single* ions in the electrolyte solution at *finite* ionic strength I is of great interest for biological systems.

* Corresponding author. E-mail: p.carloni@grs-sim.de.

[†] FZ-Juelich/RWTH Aachen University.

[‡] Pacific Northwest National Laboratory.

[§] Rush University Medical Center.

[¶] SISSA, CNR-INFN-DEMOCRITOS, and Italian Institute of Technology (IIT), SISSA Unit.

^{||} Current address: German Research School for Simulation Sciences, FZ-Juelich/RWTH Aachen University, Germany.

As we know from experiments, thermodynamic properties of electrolyte solutions at moderate I (say 0.2 m) differ already from the ideal properties found at $I = 0$. Indeed, ions, like Na^+ and K^+ , differ because they are nonideal. They have even more dramatically nonideal behavior at m ionic strength.⁹ The key quantity describing the nonideal behavior of single ions in ionic solution is the difference in excess (electro-) chemical potential (μ_X^{ex} , $X = \text{Na}^+$, K^+ , and Cl^-) between solutions at finite I and those at $I = 0$. This difference, which we write as $\Delta\mu_X^{\text{ex}}$, is given by two contributions: (i) the chemical part, which accounts for the change of intermolecular interactions between the solution molecules/ions at finite I compared to that at $I = 0$;¹⁰ and (ii) the electrical part, which is due to the electrostatic potential inside the solution generated at the interface between air and any thermodynamically stable solution. This is the so-called Galvani potential.^{11,12}

The calculation and the experimental determination of $\Delta\mu_X^{\text{ex}}$ at finite I are cumbersome. In fact, in molecular simulations approaches, such as Monte Carlo or molecular dynamics, one has to apply periodic boundary conditions to mimic macroscopic solutions; in these conditions, the non-negligible contribution due to the Galvani potential must be added.^{13,14} Although this quantity is defined mathematically unambiguously, it can be calculated only in an approximate way because of the well-known limitations of sampling and force field accuracy in molecular simulations.^{15,16} In addition, approximations must be necessarily introduced in the calculations of long-range electrostatics.^{17–19} Experimentally, it is not possible to separate the contribution of an ion from that of its counterion(s) because experiments are necessarily carried out on neutral macroscopic systems. Extra thermodynamic assumptions are then necessary.^{20–23} Indirect estimates are obtained by an analysis of different salts.²⁴ Further complications might arise from deviations from ideal conditions, which are usually assumed.^{11,12} These consider the ions as point particles, independent of size and chemical types of the ions, and the solution–air interface independent of boundary conditions.²⁵ In fact, the Galvani potential is likely to depend on the size and chemical nature of the particle. This fact is important for both theoretical and experimental estimates. Next, for the latter, the Galvani potential may depend also on complex effects specific to the setups. In particular, the thermodynamic properties of the interface may depend on finite-size effects and the presence of boundaries. Finally, in some experimental setups, non-equilibrium effects might be involved if flows are too slow to equilibrate on the time scale of experiments. The last two issues would arise in molecular simulation of the same setups.

Here we investigate the variance among force fields in predictions of $\Delta\mu_X^{\text{ex}}$ of KCl and NaCl in aqueous solution as well as the dependence of the predicted properties of Cl^- ion on its Na^+ or K^+ counterions. To this end, we performed molecular dynamics simulation of the ions in solutions based on a variety of force fields commonly used in biomolecular simulations. These include the AMBER²⁶ (the newly developed), CHARMM,^{27,28} OPLS,²⁹ and Dang95³⁰ in combination with SPC/E³¹ and TIP3P³² water models.

Prior of the prediction of $\Delta\mu_X^{\text{ex}}$, we explore the domain of applicability of these force fields. This is a nontrivial issue as these potentials are commonly calibrated by fitting to quantities like ion hydration free energy at $I = 0$ or the first peak of ion–water radial distribution functions, which are not sensitive to I .³³ This means that the nonideal effects of ions at finite strength are not considered in the parametrization. Because this issue cannot be addressed by considering $\Delta\mu_X^{\text{ex}}$ for the reasons outlined above, we resort here to a comparison between the predicted and experimental values for NaCl and KCl salts, $\Delta\mu_{\text{NaCl}}^{\text{ex}}$ and $\Delta\mu_{\text{KCl}}^{\text{ex}}$. For these, the contribution from the Galvani potential vanishes.^{14,23} Therefore, the properties of the air/water interface are not involved in the evaluation of electrostatics. This makes the calculation straightforward. In addition, experimental values are available for neutral salts solutions, such as KCl and NaCl solutions.³⁴ So far, such comparison has been made with the newly developed AMBER force field and TIP3P water solutions.²⁶ It is extended here to the other force fields listed above.

Our paper is organized as follows. Section 2 reports the thermodynamic quantities of interest in this work and the computational details. Section 3.1 assesses the accuracy of the force fields by a comparison of calculated and experimental values for $\Delta\mu_{\text{NaCl}}^{\text{ex}}$ and $\Delta\mu_{\text{KCl}}^{\text{ex}}$. Section 3.2 reports our estimate of $\Delta\mu_X^{\text{ex}}$ ($X = \text{Na}^+$, K^+ , and Cl^-), while Section 3.3 reports the calculated electrical contributions to $\Delta\mu_{\text{Na}^+}^{\text{ex}}$ and $\Delta\mu_{\text{K}^+}^{\text{ex}}$, for which corresponding values obtained by higher level calculations are available. Section 3.4 describes the dependence of the chemical contribution to $\Delta\mu_{\text{Cl}^-}^{\text{ex}}$ from the type of counterion. Section 4 discusses the implications of our results for biological systems. Section 5 summarizes the results.

2. Theory and Methods

2.1. Definition of Excess (Electro-)Chemical Potential Difference $\Delta\mu_X^{\text{ex}}$. The (electro-)chemical potential of a monovalent ion X at finite I , μ_X^I , can be expressed as^{23,35}

$$\mu_X^I = \mu_X^\circ + RT \ln \frac{I}{I^\circ} + RT \ln \gamma_X + zF\varphi^I \quad (1)$$

The reference chemical potential μ_X° is defined as the chemical potential of the X ion (e.g., Na^+) in an infinitely diluted solution (i.e., its ionic strength $I^\circ \rightarrow 0$) of one of its salts (e.g., NaCl) at room temperature and 1 atm pressure.

The activity coefficient of X is γ_X . It characterizes the nonideal thermodynamic behavior of ions due to ion–ion and ion–water interactions at at finite I . In the reference state, γ_X is assumed to be 1. $RT \ln \gamma_X$ is usually referred to as the chemical contribution to μ_X^I .

The Galvani potential at finite I is φ^I . It arises by bringing an ion from an infinite distance into the interior of the liquid phase.¹¹ The charge number is z (e.g., $z = 1$ for Na^+). While $zF\varphi^I$ includes two parts: (i) the contribution of the Volta potential, which vanishes if the solution bears no net charge (as in our case);²³ and (ii) the contribution due to the surface potential generated by the specific dipole orientation of water molecules and their quadrupole moments at the solution interface.^{36–38} This provides a non-negligible contribution to μ_X^I .^{14,23}

The excess (electro-)chemical potential which accounts for the intermolecular interaction between solution molecule/ions, is defined as¹⁰

$$\mu_X^{l,ex} = \mu_X^{o,ex} + RT \ln \gamma_X + zF(\varphi^l - \varphi^o) \quad (2)$$

$\mu_X^{o,ex}$ is the excess (electro-)chemical potential of the reference state or the hydration free energy of ions, whereas φ^o is the Galvani potential of liquid water.

The excess (electro-)chemical potential difference is then given by difference between $\mu_X^{l,ex}$ and $\mu_X^{o,ex}$

$$\Delta\mu_X^{l,ex} = RT \ln \gamma_X + zF(\varphi^l - \varphi^o) \quad (3)$$

The practical calculation of $zF(\varphi^l - \varphi^o)$ poses some challenges. It is presented in the next section, along with the straightforward calculation of $RT \ln \gamma_X$.

The excess (electro-)chemical potential of a salt (e.g., NaCl) is easily obtained from the arithmetic average of the contributions from cations and anions:

$$\Delta\mu_{NaCl}^{l,ex} = (\Delta\mu_{Na^+}^{l,ex} + \Delta\mu_{Cl^-}^{l,ex})/2 \quad (4)$$

Notice that the contribution due to the Galvani potential to $\Delta\mu_{NaCl}^{l,ex}$ and to $\Delta\mu_{KCl}^{l,ex}$ is zero because the electrolyte itself is neutral, even though its component ions are not. In fact $zF(\varphi^l - \varphi^o)$ of Na^+ (or K^+) has the opposite sign of $zF(\varphi^l - \varphi^o)$ of Cl^- .

2.2. Calculation of the Chemical Contribution to $\Delta\mu_X^{l,ex}$. $RT \ln \gamma_X$ has been calculated here from the well-known thermodynamic integration (TI) approach^{39–41} and its replica-exchange variant.^{42–44}

In the TI approach, the Hamiltonian of our initial systems (e.g., the NaCl or KCl solutions at a given ionic strength I) is gradually perturbed by inserting an ion X, and the free energy difference between the initial and final systems is then calculated. The perturbation is commonly divided into smaller windows by varying the coupling parameter λ from 0 to 1 in the Hamiltonian. $RT \ln \gamma_X$ is then obtained by numerical integration of each λ window.

$$RT \ln \gamma_X = -\frac{1}{\beta} \ln \int_0^1 d\lambda \langle U^l \rangle_{l,\lambda} + \frac{1}{\beta} \ln \int_0^1 d\lambda \langle U^o \rangle_{o,\lambda} \quad (5)$$

Here, U is the binding energy of the ion with the initial system. $\langle U \rangle_\lambda$ is the ensemble average of the thermodynamic force in each λ window.

As expected,^{14,26} the calculation of $\int_0^1 d\lambda \langle U^o \rangle_{o,\lambda}$ converges very well, and ~ 1 ns of dynamics was indeed sufficient to obtain excellent convergence. Instead, the calculation of $\int_0^1 d\lambda \langle U^l \rangle_{l,\lambda}$ turned out not to converge on the same time scale. This slow convergence may be caused by many reasons, including the fact that ion pairing is nonzero at finite I ⁴⁵ and that the diffusion of ions is slower at finite I .^{46,47} Thus, starting with different initial locations of the ion may give different results. Because of these difficulties in convergence and stability of simulations, we adopted the replica-exchange variant of TI.^{42–44} This is expected to converge much more efficiently.^{43,44} In fact, this was the case here (see Supporting Information).

2.3. Calculation of the Electrical Contribution to $\Delta\mu_X^{l,ex}$. In molecular simulations with periodic boundary conditions, the air–liquid interface is absent. The contribution $zF(\varphi^l - \varphi^o)$ due to this interface potential is expected to be significant^{48,49} and must be added. The magnitude of the interface potential depends on the details of the way long-range electrostatic calculations are calculated^{13,50} In the conditions used here (P-sum or particle-based PME),⁵¹ the interface potential can be estimated by molecular dynamics simulations of a liquid slab with vacuum interface^{52–54} (See Section 2.5 for details).

2.4. Finite Size Correction to $\Delta\mu_X^{l,ex}$. Additionally, one should consider the finite size correction on the electrostatic energy to the free energy calculations:^{17–19}

$$\frac{1}{2} q^2 \left(1 - \frac{1}{\epsilon(0)} \right) \xi_{Ew} \quad (6)$$

where q is the testing ion charge, $\epsilon(0)$ is the static dielectric constant, and $\xi_{Ew} = -2.837297/L^3$, which comes from the Madelung constant for a simple cubic lattice. This correction is expected to be much smaller than the previous one for aqueous solutions.⁵⁵ Indeed, for our box size (about 6 nm, see the next section, Section 2.5), it is expected to be 0.5 kJ/mol or smaller.⁵⁶

2.5. Computational Details. All classical molecular dynamics simulations were performed using the GROMACS package.^{57,58} Parameters and references are listed in Table 1.

Simulations were performed at the following ionic strength: 0.01, 0.15, 0.67, 1.39, 3.27, 4.28, and 4.80 m for KCl aqueous solution and 0.01, 0.15, 0.67, 1.39, 3.27, 4.80, and 5.56 m for the NaCl aqueous solution. The composition of the systems is listed in Table 2. An edge of 6.0 nm was chosen for the initial (cubic) simulation cell. The cell proved to be large enough to yield good statistics for ion pairs at low-ionic strength and to correct estimates of the bulk properties of water, such as the dielectric constant⁶¹ (also see Supporting Information). Ions were randomly placed inside a water box with separation longer than 0.45 nm. Each system was equilibrated for 1 ns with a time step of 2 fs in a No se–Hoover thermostat^{62,63} at 298 K and with a Parrinello–Rahman barostat⁶⁴ at 1 bar. The PME method⁵¹ was used to treat the long-range electrostatic interaction in the periodic system. Medium-high accuracy settings for PME were adopted,⁶⁵ in which the number of grid points for the reciprocal space calculation of the electrostatic energy calculation was 0.01 nm, a sixth degree B-spline interpolation was used, and the width of the screening Gaussian charge η was set to be 3.4 nm^{-1} . The van der Waals and short-range Coulomb interaction cutoff was 0.1 nm. The dispersion correction term was applied to the energy and pressure.⁶⁶ The SETTLE algorithm⁶⁷ was used for the rigid water models (namely TIP3P and SPC/E).

Free energy calculations were carried out in the NVT ensemble with a No se–Hoover thermostat^{62,63} at 298 K, starting from the last frame of the equilibration run. A two-stage⁶⁹ replica-exchange TI^{42–44} was used to calculate the excess chemical potential. In the first stage, the ion was gradually neutralized, whereas in the second stage, the van

Table 1. L-J Parameters of Ion Models and the Mixing Rules

model	atom	σ (nm)	ϵ (kJ/mol)	q (e)	mixing rule
AMBER ²⁶ (SPC/E)	Na ⁺	0.21595	1.47545	1.0	Lorentz–Berthelot
	K ⁺	0.28384	1.79789	1.0	
	Cl ⁻	0.48305	0.05349	-1.0	
AMBER ²⁶ (TIP3P)	Na ⁺	0.24393	0.36585	1.0	Lorentz–Berthelot
	K ⁺	0.30380	0.81041	1.0	
	Cl ⁻	0.44776	0.14891	-1.0	
CHARMM ^{27,28}	Na ⁺	0.24299	0.19623	1.0	Lorentz–Berthelot
	K ⁺	0.31426	0.36401	1.0	
	Cl ⁻	0.40447	0.62760	-1.0	
OPLS ²⁹	Na ⁺	0.33304	0.01160	1.0	geometric
	K ⁺	0.49346	0.00137	1.0	
	Cl ⁻	0.44172	0.49283	-1.0	
Dang95 ³⁰	Na ⁺	0.25840	0.41840	1.0	Lorentz–Berthelot
	K ⁺	0.33320	0.41840	1.0	
	Cl ⁻	0.44010	0.41840	-1.0	
SPC/E ³¹	O	0.31660	0.65060	-0.8476	
	H	0.00	0.00	0.4238	
TIP3P ³²	O	0.31510	0.63640	-0.834	
	H	0.00	0.00	0.417	
	H ^{59,60}	0.04000	0.19246	0.417	

Table 2. Numbers of Water N_{water} and Ion Pairs $N_{\text{ion pair}}$ in the Simulation System

	ionic strength (m)							
	0.01	0.15	0.67	1.39	3.27	4.28	4.80	5.56
$N_{\text{water}}^{\text{68}}$	7804	7764	7624	7436	6986	6766	6656	6504
$N_{\text{ion pair}}$	0	20	90	184	409	519	574	650

der Waals interaction was slowly switched off. A soft-core potential was used to avoid singularity of force when testing whether an ion appeared or disappeared.⁷⁰ At each stage, 10 equispaced λ windows were sampled. For each λ window, simulations were started from uncorrelated configurations. Exchanges between neighboring λ configurations were attempted every 3 ps. The first ps of each of these 3 ps simulations was discarded. A total of 2 ns long trajectories were collected for each replica-exchange TI stage. The trapezoid rule was used to integrate the averaged thermodynamics force profile. The statistical error of each window was estimated by block averaging,⁷¹ and the final error of the free energy difference was calculated by error propagation.

The calculation of the surface potential was carried out in an orthorhombic cell in a 8.4 nm thick slab containing water and ions in the same composition as used in the free energy calculation. The spacing along the z -axis was large enough to create two vapor–liquid interfaces and three-dimensional (3D) periodic boundary conditions were applied. The box size was chosen around $2.8 \times 2.8 \times 8.4$ nm, as is usual in simulations of the surface potential of an air–liquid interface.^{48,49,52–54} Each simulation was performed for 10 ns in NVT ensemble with a Nosé–Hoover thermostat at 298K.^{62,63} Electrostatic potential was evaluated from the averaged charge density profile along the z -axis. The density was calculated on a 0.02 nm grid.⁷²

3. Result and Discussions

3.1. $\Delta\mu_{\text{KCl}}^{\text{ex}}$ and $\Delta\mu_{\text{NaCl}}^{\text{ex}}$: Comparison between Calculated Values and Experiment. Our calculation for salts $\Delta\mu_{\text{KCl}}^{\text{ex}}$ and $\Delta\mu_{\text{NaCl}}^{\text{ex}}$ using the newly developed AMBER-TIP3P force field²⁶ reproduces quantitatively the experimental data

(Figure 1), as previously reported.^{73,74} The CHARMM-TIP3P and Dang95-SPC/E force field-based calculations predict accurately the values for the KCl and NaCl solutions, respectively (Figure 2). All the other potential models are not as good (Figure 2). It is of interest to notice that a recent study⁷⁵ showed that the CHARMM parameters for Na–Cl interactions generated from the Lorentz–Berthelot combination rule lead to a larger underestimation of osmotic pressure—a probe for ions activity¹² than the corresponding one for K–Cl interactions.

3.2. Calculation of $\Delta\mu_X^{\text{ex}}$. The calculated values for individual ions $\Delta\mu_X^{\text{ex}}$ ($X = \text{Na}^+$, K^+ , and Cl^-) are as scattered at finite I as the corresponding ones for the KCl

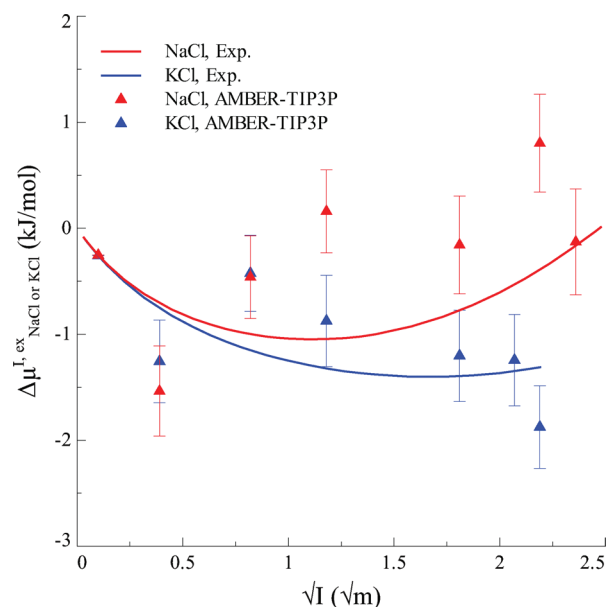


Figure 1. Calculated excess (electro-)chemical potential differences for KCl $\Delta\mu_{\text{KCl}}^{\text{ex}}$ and NaCl $\Delta\mu_{\text{NaCl}}^{\text{ex}}$, based on the newly developed AMBER-TIP3P force field,²⁶ plotted as a function of the square root of the m ionic strength. Comparing is made with experimental data.³⁴

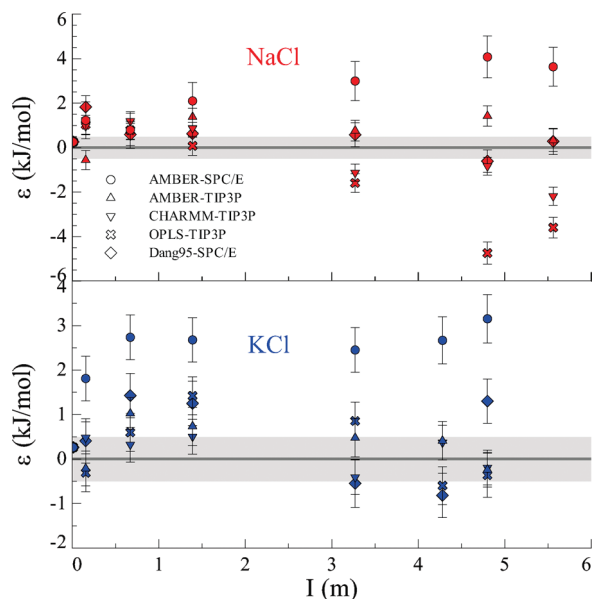


Figure 2. Deviations ε of calculated excess (electro-)chemical potential differences for KCl $\Delta\mu_{\text{KCl}}^{\text{ex}}$ and NaCl $\Delta\mu_{\text{NaCl}}^{\text{ex}}$ from experimental data³⁴ plotted as a function of the m ionic strength. The shadow area covers the deviation ε within ± 0.5 kJ/mol. The results obtained with all the force fields considered in this work are presented.

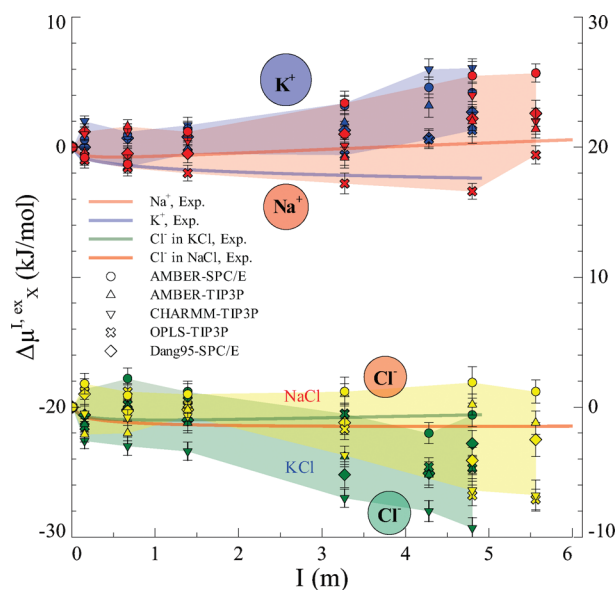


Figure 3. Calculated excess (electro-)chemical potential differences for single ions $\Delta\mu_X^{\text{ex}}$ ($X = \text{Na}^+$, K^+ , and Cl^-) in KCl and NaCl solutions, plotted as a function of the m ionic strength. The results obtained with all the force fields considered in this work are presented. Experimentally derived estimates are also reported.²⁴

and NaCl salts (Figure 3). This hints that thermodynamics of ions using different force fields differ from each other at finite I .

The magnitude of these values for $\Delta\mu_X^{\text{ex}}$ is comparable with that of the available experimentally derived data.²⁴ However, the calculated $\Delta\mu_{\text{K}^+}^{\text{ex}}$ increases with I more than $\Delta\mu_{\text{Na}^+}^{\text{ex}}$. The opposite trend is found in the experimental estimates.⁷⁶ Similarly, the calculated $\Delta\mu_{\text{Cl}^-}^{\text{ex}}$ decreases with I more in the KCl solution than it does in the NaCl solution.

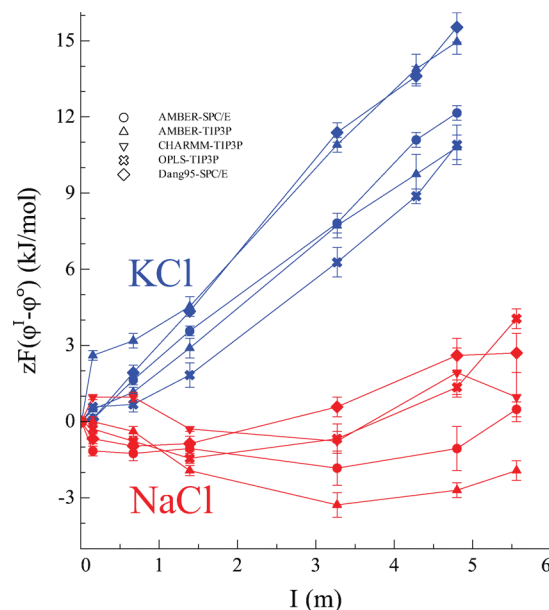


Figure 4. Calculated electrical contribution $zF(\varphi^I - \varphi^0)$ to $\Delta\mu_X^{\text{ex}}$ for K^+ and Na^+ in KCl and NaCl, respectively ($z = 1$), plotted as a function of the m ionic strength I . The results obtained with all the force fields considered in this work are presented.

The opposite occurs for the experimentally derived values. These significant discrepancies may arise from several errors and assumptions from both theory and experiments, as discussed in the Introduction Section.

To provide some hints of the origin of errors specific to the calculations, we focus here on comparisons against results obtained using higher level calculations. These are available only for the electrical contribution $zF(\varphi^I - \varphi^0)$.

3.3. Some Considerations on the Electrical Contribution $zF(\varphi^I - \varphi^0)$. In this section, we report our calculated values for $zF(\varphi^I - \varphi^0)$ at finite I and compare with previous calculations, based on polarizable force fields.^{48,49} Notice that also the latter results, even though they are expected to be much more accurate than those based on a nonpolarizable force field, still cannot present the exact Galvani potential. This is because they do not fully take into account the contribution due to the molecular quadrupoles.^{36,37}

The calculated electrical contribution $zF(\varphi^I - \varphi^0)$ to $\Delta\mu_{\text{K}^+}^{\text{ex}}$ increases linearly with I for all the force fields used here, ranging from 0 to 16 kJ/mol (Figure 4).^{77,78} The range of the calculated values of $zF(\varphi^I - \varphi^0)$ is comparable to that obtained by polarizable ion/water force field-based calculations at $I = 1$ m (from 1 to 4 kJ/mol versus 3.4 kJ/mol).^{48,79}

The overall values of calculated $zF(\varphi^I - \varphi^0)$ for a Na^+ range is from -3 to 3 kJ/mol. Thus, the values of $zF(\varphi^I - \varphi^0)$ at $I = 1$ m range from -1 to 0.5 kJ/mol, to be compared with the value obtained with a polarizable force field of 3.5 kJ/mol.^{49,79} We conclude that nonpolarizable models for the NaCl solution are not able to reproduce the results of polarizable models.

The experiments estimated an increase of the Galvani potential in both KCl and NaCl electrolyte solutions at finite I .^{37,80,81} However the quantities are all much smaller (about

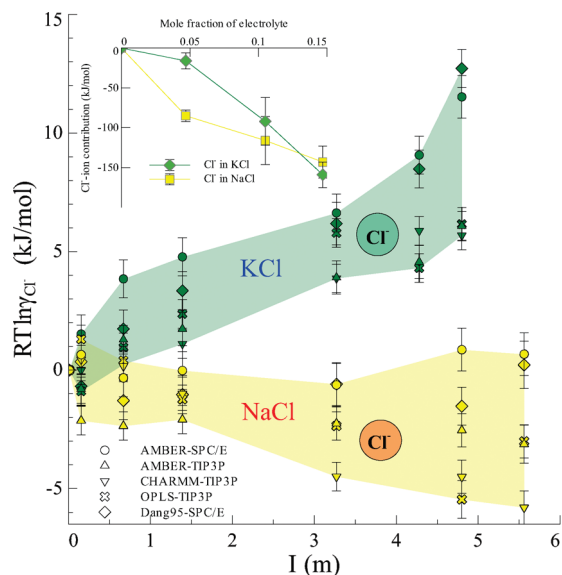


Figure 5. Calculated chemical contribution $RT \ln \gamma_{\text{Cl}^-}$ in KCl and NaCl aqueous solutions, plotted as a function of the m ionic strength. The results obtained with all the force fields considered in this work are presented. Inset: Cl^- -ion electrostatic contribution to $RT \ln \gamma_{\text{Cl}^-}$ based on the newly developed AMBER-SPC/E force field.²⁶

0.2 and 0.3 kJ/mol for KCl^{80,82} and NaCl, respectively, at $I = 1$ m).^{80,82} The very large discrepancies between theory and experiment reflect the difficulties in experimental measurement of the Galvani potential (see Introduction Section) as well as limitations of the molecular simulation methods outlined in the Introduction.

3.4. $RT \ln \gamma_{\text{Cl}^-}$: Dependence from the Types of Counterions. The chemical contribution $RT \ln \gamma_{\text{Cl}^-}$ as a function of I depends on the type of counterion for all the force fields used here (Figure 5).

As mentioned before, $RT \ln \gamma_{\text{Cl}^-}$ reflects the change of intermolecular interactions between Cl^- -ion and Cl^- -water at finite I . This change in electrolyte solution is often attributed to the electrostatic interactions as a first approximation.⁸³ We find the Cl^- -ion electrostatic contribution to $RT \ln \gamma_{\text{X}}$ of the NaCl solution is dramatically different from that of the KCl solution, obtained from a calculation based on the newly developed AMBER-SPC/E force field^{26,84} (inset in Figure 5). Similar conclusions can be drawn for Cl^- -water electrostatic contributions in the two salt solutions (data not shown).

4. Implication for Biological Systems

The success of predicting the values for salts is gratifying with some of the force fields considered here, especially considering their very simple functional form. The success testifies to the care with which force fields have been developed. However, the challenges reported previously,^{13,20–22,55,85} and addressed here, do remain in the prediction of $\Delta\mu_{\text{X}}^{\text{I,ex}}$ ($\text{X} = \text{Na}^+$, K^+ , and Cl^-) and in particular of the electric contribution to it (see Sections 2.3 and 3.3). These difficulties may be even larger when modeling biological systems. Such difficulties do not come without consequence. Consider the simple identification of an ion channel, as done by (literally) thousands of laboratories

every day. That identification depends on the measurement and identity of the (so-called) reversal potential,^{86,87} which is the experimental estimator of the gradient of chemical potential or the equilibrium potential, as it was called by Hodgkin and Huxley.^{88,89} The name of the channel is often determined by its selectivity^{90–93} (e.g., sodium, potassium, or chloride channels), and that in turn depends on the identification of the reversal potential with the gradient of chemical potential of one ion. If in fact $\Delta\mu_{\text{X}}^{\text{I,ex}}$ is not accurately included^{94–98} in the calculation of the gradient of chemical potential (when using concentration of ions as inputs), then the channel identification may be askew.⁹⁷

The selectivity properties of ion channels are crucially important to their function. Ions that differ in their nonideal properties, like Na^+ and K^+ , carry different ‘messages’ (i.e., signals) to different systems of the cell, and so there is enormous literature trying to measure, understand, simulate, control, and even synthesize^{99–101} the selectivity of different types of channels. Estimates and computations of selectivity depend critically on estimates of $\Delta\mu_{\text{X}}^{\text{I,ex}}$, because many types of ions differ only because they are nonideal. Similar considerations^{87,102–113} are likely to apply to a myriad of other biological events. Many important biological properties arise because of the nonideal properties of individual types of ions.

5. Conclusion

We have established the quality of a variety of standard ion/water force fields commonly used in biological simulation, for the calculation of the excess (electro-)chemical potential for KCl $\Delta\mu_{\text{KCl}}^{\text{I,ex}}$ and for NaCl $\Delta\mu_{\text{NaCl}}^{\text{I,ex}}$. Specifically, the AMBER²⁶ (the newly developed), CHARMM,^{27,28} OPLS,²⁹ and Dang95³⁰ were considered in combination with SPC/E³¹ and TIP3P³² water models. The calculation based on the newly developed AMBER-TIP3P agrees well with the experimental values for both KCl and NaCl solutions, as previously reported.⁷³ Instead the CHARMM-TIP3P potential agrees well with the KCl salt, whereas the Dang95-SPC/E potential agrees well with the NaCl salt. The other potential models do not give good results for either of the two aqueous solutions studied. Hence, care should be taken in biomolecular simulations when using these force fields at physiological I .

The calculated $\Delta\mu_{\text{Na}^+}^{\text{I,ex}}$ values are similar to those of $\Delta\mu_{\text{K}^+}^{\text{I,ex}}$. The calculated values are as scattered at finite I as the corresponding ones for the KCl and NaCl salts. Only the calculated electric contribution $zF(\varphi^I - \varphi^0)$ of K^+ is consistent with reported higher level calculations with polarizable ion/water force fields.⁴⁸

The calculated chemical contribution $RT \ln \gamma_{\text{Cl}^-}$ to $\Delta\mu_{\text{Cl}^-}^{\text{I,ex}}$ depends on the type of counterions present. This result may be of interest for force field calculations of Cl^- -dependent biological systems (such as chloride channels).¹¹⁴

Acknowledgment. The author (C. Z.) thanks F. Marinelli for helpful discussion on the replica-exchange method. We thank the reviewers for their highly valuable comments on the manuscript.

Supporting Information Available: Tests of the convergence of free-energy calculation, calculated dielectric constants as a function of ionic strength, estimates of the

Galvani potential of pure water as well as the density profiles of concentrated salt solutions. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) "m" is molal scale not molar scale, i.e., m is mol/1 kg of H₂O.
- (2) Costanzo, L. S. *Physiology*; Elsevier/Saunders: Philadelphia, PA, 2006; pp 2–4.
- (3) Gouaux, E.; MacKinnon, R. *Science* **2005**, *310*, 1461.
- (4) Doyle, D. A.; Cabral, J. M.; Pfuetzner, R. A.; Kuo, A.; Gulbis, J. M.; Cohen, S. L.; Chait, B. T.; MacKinnon, R. *Science* **1998**, *280*, 69.
- (5) Domene, C.; Vemparala, S.; Furini, S.; Sharp, K.; Klein, M. L. *J. Am. Chem. Soc.* **2008**, *130*, 11.
- (6) Jayaram, B.; Jain, T. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 343.
- (7) Auffinger, P.; Hashem, Y. *Curr. Opin. Struct. Biol.* **2007**, *17*, 325.
- (8) Chu, V. B.; Bai, Y.; Lipfert, J.; Herschlag, D.; Doniach, S. *Curr. Opin. Chem. Biol.* **2008**, *12*, 619.
- (9) Lee, L. L. *Molecular Thermodynamics of Electrolyte Solutions*; World Scientific: New York, 2008; pp 11–38.
- (10) Beck, T. L.; Paulaitis, M. E.; Pratt, L. R. *The Potential Distribution Theorem and Models of Molecular Solutions*; Cambridge University Press: Cambridge, U.K., 2006; pp 45–68.
- (11) Butt, H. J.; Graf, K.; Kappl, M. *Physics and Chemistry of Interfaces*; Wiley-VCH: Weinheim, Germany, 2006; pp 77–79.
- (12) *Liquids, Solutions, and Interfaces: From Classical Macroscopic Descriptions to Modern Microscopic Details*; Fawcett, W. R., Ed.; Oxford University Press: New York, 2004; pp 3–147.
- (13) Kastenholz, M. A.; Hünenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 224501.
- (14) Lamoureux, G.; Roux, B. *J. Phys. Chem. B* **2006**, *110*, 3308.
- (15) van Gunsteren, W. F.; et al. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064.
- (16) McDowell, S. E.; Spackova, N.; Sponer, J.; Walter, N. G. *Biopolymer* **2007**, *82*, 169.
- (17) Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem.* **1996**, *100*, 1206.
- (18) Hummer, G.; Pratt, L. R.; García, A. E. *J. Chem. Phys.* **1997**, *107*, 9275.
- (19) Hünenberger, P. H.; McCammon, J. A. *J. Chem. Phys.* **1999**, *110*, 1856.
- (20) Conway, B. E. *J. Sol. Chem.* **1978**, *7*, 720–770.
- (21) Marcus, Y. *J. Chem. Soc. Faraday Trans.* **1987**, *83*, 2985.
- (22) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V. *J. Phys. Chem. A* **1998**, *102*, 7787.
- (23) Fawcett, W. R. *Langmuir* **2008**, *24*, 9868.
- (24) Wilczek-Vera, G.; Rodil, E.; Vera, J. H. *AIChE J.* **2004**, *50*, 445.
- (25) Within these simplifying assumptions, the Galvani potential is the same for K⁺ and Na⁺, and it is opposite in sign for Cl⁻.
- (26) Joung, I. S.; Cheatham, T. E., III *J. Phys. Chem. B* **2008**, *112*, 9020.
- (27) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050.
- (28) Roux, B. *Biophys. J.* **1996**, *71*, 3177.
- (29) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.
- (30) Dang, L. X. *J. Am. Chem. Soc.* **1995**, *117*, 6954.
- (31) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.
- (32) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (33) Patra, M.; Karttunen, M. *J. Comput. Chem.* **2004**, *25*, 678.
- (34) Hamer, W. J.; Wu, Y.-C. *J. Phys. Chem. Ref. Data* **1972**, *1*, 1047.
- (35) Fawcett, W. R. *Liquids, Solutions, and Interfaces: From Classical Macroscopic Descriptions to Modern Microscopic Details*; Oxford University Press: New York, 2004; pp 395–422.
- (36) Paluch, M. *Adv. Colloid Interface Sci.* **2000**, *84*, 27.
- (37) Petersen, P. B.; Saykally, R. J. *Annu. Rev. Phys. Chem.* **2006**, *57*, 333.
- (38) Continuum models encounter difficulties in describing the Galvani potential because the latter depends on the polarization of the system.
- (39) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.
- (40) Ferrario, M.; Ciccoti, G.; Spohr, E.; Cartailier, T.; Turq, P. *J. Chem. Phys.* **2002**, *117*, 4947.
- (41) *Free Energy Calculations*; Chipot, C., Pohorille, A., Eds.; Springer-Verlag: Berlin, Germany, 2007; pp 33–72.
- (42) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058.
- (43) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem B* **2003**, *107*, 13711.
- (44) Jiang, W.; Hodoscek, M.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 2583.
- (45) Uchida, H.; Matsuoka, M. *Fluid Phase Equilib.* **2004**, *219*, 49.
- (46) Chowdhuri, S.; Chandra, A. *J. Chem. Phys.* **2001**, *115*, 3732.
- (47) Instead, the diffusion of ions has no impact on the convergence of TI calculation at infinite dilution.
- (48) Wick, C. D.; Dang, L. X.; Jungwirth, P. *J. Chem. Phys.* **2006**, *125*, 024706.
- (49) Bauer, B. A.; Patel, S. *J. Chem. Phys.* **2010**, *132*, 024713.
- (50) Kastenholz, M. A.; Hünenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 124106.
- (51) Sagul, C.; Darden, T. A. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155.
- (52) Feller, S. E.; Pastor, R. W.; Rojnuckarin, A.; Bogusz, S.; Brooks, B. R. *J. Phys. Chem.* **1996**, *100*, 17011.
- (53) Sokhan, V. P.; Tildesley, D. J. *Mol. Phys.* **1997**, *92*, 625.
- (54) Lamoureux, G.; Mackerell, A. D., Jr.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185.

- (55) Cheng, J.; Sulpizi, M.; Sprik, M. *J. Chem. Phys.* **2009**, *131*, 154504.
- (56) The calculated static dielectric constants $\epsilon(0)$ of KCl and NaCl solutions comparing with experimental data as a function of the molal ionic strength can be seen in Supporting Information, Figure S2.
- (57) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701.
- (58) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
- (59) MacKerell, A. D., Jr.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586.
- (60) Modified TIP3P water in CHARMM⁵⁹.
- (61) Lin, Y.; Baumketner, A.; Deng, S.; Xu, Z.; Jacobs, D.; Cai, W. *J. Chem. Phys.* **2009**, *131*, 154103.
- (62) Noše, S. *J. Chem. Phys.* **1984**, *81*, 511.
- (63) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695.
- (64) Parrinello, M.; Rahman, A. *Phys. Rev. Lett.* **1980**, *45*, 1196.
- (65) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.
- (66) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publications: Oxford, U.K., 1987; pp 64–68.
- (67) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952.
- (68) Exact number depends on the water model and the salt type.
- (69) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395.
- (70) Beutler, T. C.; Mark, A. E.; Van Schaik, R. C.; Gerber, P. R.; Van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529.
- (71) Hess, B. *J. Chem. Phys.* **2002**, *116*, 209.
- (72) Note that it is also possible to obtain the Galvani potential by creating a virtual air–solution interface with those snapshots from simulations of bulk solutions under PBC and then integrating the charge density.
- (73) Joung, I. S.; Cheatham, T. E. *J. Phys. Chem. B* **2009**, *113*, 13279.
- (74) Notice that the time-scale of our simulation is shorter than that of these authors.⁷³ They use straightforward TI instead of replica-exchange TI. The latter converges faster, See Figure S1 in Supporting Information.
- (75) Luo, Y.; Roux, B. *J. Phys. Chem. Lett.* **2010**, *1*, 183.
- (76) Similar trends were also founded experimentally²⁴ in the presence of an anion other than Cl⁻.
- (77) Table S1 in Supporting Information presents a comparison of φ° values, which is not crucial for the $\Delta\mu_X^{\text{ex}}$ but may be relevant as a reference.
- (78) We only observed a slight preference of the anions at the interface than cations in the simulations (See Figure S3 in Supporting Information).
- (79) The actual values reported in refs 48 and 49 are $(\varphi^I - \varphi^\circ)$. For the sake of clarity, here we report $zF(\varphi^I - \varphi^\circ)$, which is the quantity of interest here.
- (80) Randles, J. E. *Phys. Chem. Liq.* **1977**, *7*, 107.
- (81) Jungwirth, P.; Tobias, D. J. *Chem. Rev.* **2006**, *106*, 1259.
- (82) Jarvis, N. J.; Scheiman, M. A. *J. Phys. Chem.* **1968**, *72*, 74.
- (83) Wright, M. R. *An Introduction to Aqueous Electrolyte Solutions*; Wiley: Chichester, U.K., 2007.
- (84) We expect similar results for all the other force fields as they have the same trend in Figure 5. However, free energy decomposition is force field and path dependent.
- (85) Harder, E.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 234706.
- (86) Hille, B. *Ionic Channels of Excitable Membranes*, 3rd ed.; Sinauer Associates Inc.: Sunderland, MA, 2001; pp 1–19.
- (87) Zuhlke, R. D.; Pitt, G. S.; Deisseroth, K.; Tsien, R. W.; Reuter, H. *Nature* **1999**, *399*, 159.
- (88) Hodgkin, A.; Huxley, A.; Katz, B. *Arch. Sci. Physiol.* **1949**, *3*, 129.
- (89) Hodgkin, A. *J. Physiol.* **1976**, *263*, 1.
- (90) Conley, E. C. *The Ion Channel Facts Book. I. Extracellular Ligand-gated Channels*; Academic Press: New York, 1996; pp 3–11.
- (91) Conley, E. C. *The Ion Channel Facts Book. II. Intracellular Ligand-gated Channels*; Academic Press: New York, 1996; pp 3–20.
- (92) Conley, E. C.; Brammar, W. *The Ion Channel Facts Book III: Inward Rectifier and Intercellular Channels*; Academic Press: New York 2000; pp 3–21.
- (93) Conley, E. C.; Brammar, W. *The Ion Channel Facts Book IV: Voltage Gated Channels*; Academic Press: New York, 1999; pp 3–21.
- (94) Barry, P. H. *Am. J. Physiol.* **1990**, *259*, S15.
- (95) Barry, P. H. *Ann. Biomed. Eng.* **1994**, *22*, 218.
- (96) Barry, P. H. *J. Neurosci. Methods* **1994**, *51*, 107.
- (97) Barry, P. H. *Cell Biochem. Biophys.* **2006**, *46*, 143.
- (98) Ng, B.; Barry, P. H. *J. Neurosci. Methods* **1995**, *56*, 37.
- (99) Miedema, H.; Meter-Arkema, A.; Wierregna, J.; Tang, J.; Eisenberg, B.; Nonner, W.; Hektor, H.; Gillespie, D.; Meijberg, W. *Biophys. J.* **2004**, *87*, 3137.
- (100) Miedema, H.; Vrouenraets, M.; Wierregna, J.; Eisenberg, B.; Gillespie, D.; Meijberg, W.; Nonner, W. *Biophys. J.* **2006**, *91*, 4392.
- (101) Vrouenraets, M.; Wierregna, J.; Meijberg, W.; Miedema, H. *Biophys. J.* **2006**, *90*, 1202.
- (102) Berg, J. M. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 405.
- (103) Berg, J. M. *J. Biol. Chem.* **1990**, *265*, 6513.
- (104) Cantwell, M. A.; Di Cera, E. *J. Biol. Chem.* **2000**, *275*, 39827.
- (105) Berg, J. M.; Godwin, H. A. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 357.
- (106) Carnell, C. J.; Bush, L. A.; Mathews, F. S.; Di Cera, E. *Biophys. Chem.* **2006**, *121*, 177.
- (107) De Gristofaro, R.; Fenton II, J. W.; Di Cera, E. *J. Mol. Biol.* **1992**, *226*, 263.
- (108) Di Cera, E. *Biopolymer* **1994**, *34*, 1001.
- (109) Doroshenko, P. A.; Kostyuk, P. G.; Lukyanetz, E. A. *Neurosci.* **1998**, *27*, 1073.
- (110) Eisenberg, R. S. *J. Membr. Biol.* **1990**, *115*, 1.

- (111) Lambers, T. T.; Mahieu, F.; Oancea, E.; Hoofd, L.; de Lange, F.; Mensenkamp, A. R.; Voets, T.; Nilinus, B.; Clapham, D. E.; Hoenderop, J. G.; Bindels, R. J. *EMBO J.* **2006**, 25, 2978.
- (112) Tripathy, A.; Xu, L.; Mann, G.; Meissner, G. *Biophys. J.* **1995**, 69, 106.
- (113) Vescovi, E. G.; Ayala, Y. M.; Di Cera, E.; Groisman, E. A. *J. Biol. Chem.* **1997**, 272, 1440.
- (114) Suzuki, M.; Morita, T.; Iwamoto, T. *Cell. Mol. Life Sci.* **2006**, 63, 12.

CT9006579

Starting-Condition Dependence of Order Parameters Derived from Molecular Dynamics Simulations

Samuel Genheden,[†] Carl Diehl,[‡] Mikael Akke,[‡] and Ulf Ryde^{*,†}

Department of Theoretical Chemistry, Lund University, Chemical Centre, P.O. Box 124, SE-221 00 Lund, Sweden and Center for Molecular Protein Science, Biophysical Chemistry, Lund University, P.O. Box 124, SE-221 00 Lund, Sweden

Received December 28, 2009

Abstract: We have studied how backbone N–H S^2 order parameters calculated from molecular dynamics simulations depend on the method used to calculate them, the starting conditions, and the length of the simulations. Using the carbohydrate binding domain of galectin-3 in the free and lactose-bound states as a test case, we compared the calculated order parameters with experimental data from NMR relaxation. The results indicate that the sampling can be improved by using several starting structures, taking into account conformational heterogeneity reported in crystal structures. However, the improvement is rather limited, and for 93% of the dihedrals that have alternative conformations in the crystal structures, the conformational space is well sampled even if a single conformation is used as the starting structure. Moreover, the agreement with experimental data is improved when using several short simulations, rather than a single long simulation. In the present case, we find that ~ 10 independent simulations provide sufficient sampling, and the ideal length of the simulations is ~ 10 ns, which is $\sim 25\%$ longer than the global correlation time for rotational diffusion. On the other hand, the equilibration time appears to be less important, and our results suggest that an equilibration time of 0.25 ns is sufficient. We have also compared four different methods to extract the order parameters from the simulations, namely, the autocorrelation function and isotropic reorientational eigenmode dynamics using three different window sizes. Overall, the four methods yield comparable results, but large differences between the methods may serve to pinpoint cases for which the calculated parameters are unreliable.

Introduction

Nuclear spin relaxation is a powerful experimental technique that provides site-specific information on dynamics and conformational entropy.^{1–3} Such measurements are normally interpreted in the context of the model-free approach,^{4–6} yielding a generalized order parameter (S^2) for each studied bond vector. Typically, NMR spectroscopic investigations of conformational dynamics focus on a relatively limited subset of bond vectors, although continuous method development aims to expand this set.^{7–11} Thus, investigations based

solely on experimental data inevitably undersample the conformational entropy of the system, although recent results suggest that order parameters for selected subsets of bond vectors actually capture conformational entropy quite well.¹²

Therefore, NMR spin relaxation experiments can favorably be combined with molecular dynamics (MD) simulations to augment the information content of experimental order parameters.^{13,14} MD simulations provide a detailed picture of the motions of all atoms considered, with an accuracy and precision similar to that of NMR experiments.¹⁵ Thus, MD simulations offer a route to interpret in greater detail the results from spin relaxation experiments, provided that the two techniques yield commensurate results.¹⁶ In particular, MD simulations can provide the probability distribution of the conformational substates, including those degrees of

* Corresponding author. Tel: +46 - 46 2224502. Fax: +46 - 46 2228648. E-mail: Ulf.Ryde@teokem.lu.se.

[†] Department of Theoretical Chemistry.

[‡] Center for Molecular Protein Science.

freedom that are not probed by spin relaxation measurements. Once an MD-generated conformational ensemble has been validated by experimental NMR data, it is therefore possible to calculate the total conformational entropy of the system and to address other issues such as the degree of coupling between bond vector motions. In addition, MD simulations offer a high-resolution view of the motional mechanisms that cannot be determined directly from the NMR relaxation data. Conversely, comparisons of order parameters obtained with MD and NMR have frequently been used to judge and improve the quality of MD force fields.^{13,17–19}

A major issue for the calculation of generalized S^2 order parameters²⁰ from MD simulations is the convergence—typically quite long simulations are needed to reach convergence. Related to this issue is how the results depend on the starting conditions of the MD simulations. Several studies have shown that results of MD simulations, e.g., order parameters, strongly depend on the starting structure.^{21–23} It has been much discussed whether it is more favorable to run a single long simulation or several shorter ones.^{24–26}

In this paper, we examine a related problem: Many crystal structures, especially those obtained at a high resolution, show residues with multiple conformations. This provides a practical problem for MD simulations, because only a single structure is normally treated in the simulations. Which of these conformations should be selected as the starting structure, and how do the results depend on this selection? Is it necessary to start from many different conformations to cover the conformational space appropriately? Can we use this information to speed up the convergence of calculated properties? In this paper, we provide a systematic investigation of these issues. In particular, we study how the calculated order parameters vary and how they compare to experimental NMR data.

As a model system, we have studied the carbohydrate-recognition domain of galectin-3 (Gal3), for which high-resolution X-ray structures are available.²⁷ Galectins represent a family of proteins that preferentially bind β -galactoside-containing glycans composed of N-acetylglucosamine.^{28,29} They are involved in a wide variety of extracellular and intracellular processes, e.g., cancer,^{30,31} immunity, inflammation,^{32,33} and RNA splicing.^{34,35} The Gal3 structure consists of two antiparallel β sheets of six and five strands (Figure 1).^{27,28} The saccharide-binding site is defined by a shallow groove formed by the six-stranded β sheet and surrounding loops. Galectin–monosaccharide interactions are relatively weak, with dissociation constants on the order of 0.1–1 mM. The binding free energy is in general dominated by enthalpic contributions and has a minor unfavorable entropic contribution.³⁶ Typically, two to five hydrogen bonds are formed between the carbohydrate ligand and Gal3, in addition to favorable van der Waals interactions.

Methods

MD Simulations. The carbohydrate-recognition domain of the protein galectin-3 (Gal3) was studied both in the unbound form (Gal3-apo) and in complex with lactose (Gal3-Lac). The simulations were based on an unpublished 1.08

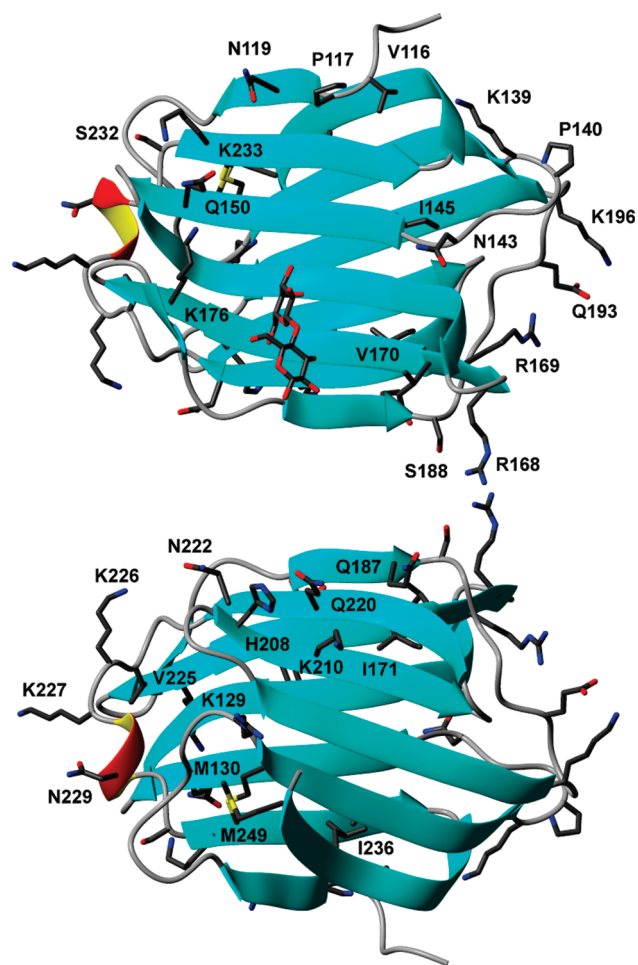


Figure 1. Structure of Gal3 with important residues indicated: (top) the side with the binding site, (bottom) the back side.

Å structure of Gal3-apo and on a 1.35 Å structure of Gal3-Lac (PDB code 2nn8).²⁷ The two structures are very similar, with a backbone RMSD of 0.22 Å. In both crystals, several residues are reported with two conformations (27 for Gal3-apo and 16 for Gal3-Lac), all with an occupancy of 0.5. In the MD simulations, we need to select one of those conformations for the starting structure, but the choice is totally arbitrary and might possibly bias the final results. To investigate the effect of the selected conformation, residues were divided into five groups of nearby residues, and the 32 possible permutations of these groups were prepared in which residues in the same group had the same conformation, A or B. The groups are specified in Tables S1 and S2 in the Supporting Information. For about a third of the residues with alternative conformations, there is a change in the hydrogen-bond pattern around that residue, as is also specified in Tables S1 and S2.

All simulations were run using the Amber 10 sander module.³⁷ The lactose molecule was described with the glycam06 force field and the protein with the Amber99SB force field.¹⁷ Protons were added with the leap module of Amber, and the protonation states were as described previously.³⁸ The systems were solvated in an octahedral box of TIP4P-Ewald waters,³⁹ extending at least 9 Å from the protein. The SHAKE algorithm⁴⁰ was used to constrain bonds involving hydrogen atoms, making a 2 fs time step

possible. The temperature was kept constant at 300 K using Langevin dynamics⁴¹ with a collision frequency of 2.0 ps⁻¹. The pressure was kept at 1 atm using a weak-coupling approach,⁴² with isotropic position rescaling and a relaxation time of 1 ps. Long-range electrostatics were treated with the particle-mesh Ewald approach⁴³ with a fourth-order B-spline interpolation and a tolerance of 10⁻⁵. The nonbonded cutoff was 8 Å, and the nonbonded pair list was updated every 50 fs.

The systems were energy minimized for 1000 steps, restraining all water molecules and heavy atoms to their start positions with a force constant of 418 kJ mol⁻¹ Å⁻². This was followed by a 20 ps equilibration with the same restraints and constant pressure, 50 ps equilibration without any restraints at constant pressure, and 200 ps equilibration at constant volume and no restraints. Finally, a 20 ns production run was performed, still at a constant volume. Coordinates were saved every 1 ps for the calculation of order parameters. On the basis of the stability of the backbone RMSD, the first 5 ns were discarded from subsequent analysis, unless otherwise stated.

We also performed 10 independent simulations of the proteins, started with all residues in the A conformation. The simulation protocol was as described above, but the production simulation was extended to 40 ns. The independent simulations were generated by using different random starting velocities.

Thus, we have run 32 simulations of 20 ns length, starting from different conformations, and 10 simulations of 40 ns length, starting from the same A conformation, but with different velocities. In the following, we will discuss the results obtained from different subsets of these simulations. These subsets will be referred to by the number of simulations, followed by the letter M for mixed conformations or A for the A conformation and then by the length of the simulation in nanoseconds, preceded by “×”. For example, 10 simulations of 5 ns length, started from different conformations will be denoted 10M×5.

MD-Derived Order Parameters. Two different methods were used to calculate order parameters from the MD simulations. In the first, order parameters were estimated from the plateau value ($\lim(t \rightarrow \infty) C_2(t)$) of the following time autocorrelation function (ACF):

$$C_2(t) = A \langle 3(\vec{\mu}(\tau)\vec{\mu}(\tau + t))^2 - 1 \rangle \quad (1)$$

where A is a constant (including the length of the N–H vector) and the average was calculated over the trajectory.⁴⁴ The unit vectors $\vec{\mu}(\tau)$ and $\vec{\mu}(\tau + t)$ describe the orientation of the N–H vector of interest at times τ and $\tau + t$ in relation to a fixed reference frame. This ACF was calculated using the Amber 10 ptraj module, and the overall tumbling was removed by fitting the backbone heavy atoms to the first snapshot. It is not fully straightforward to determine the plateau value of C_2 , because C_2 becomes noisy at large values of the time delay, t , owing to the finite sampling time. Therefore, C_2 was only calculated for t up to $\sim 1/10$ of the total simulation time.^{45–47} The order parameters were then obtained by fitting $C_2(t)$ to an exponential function of the form

$$A + B e^{-Ct} + D e^{-Et} \quad (2)$$

where A , B , C , D , and E are fitted coefficients²⁰ and the order parameter can be identified with A .^{44,13} Statistical errors of the order parameters were estimated using a bootstrap procedure on the residuals from the exponential fit, using 1000 samples.⁴⁸

Alternatively, order parameters were extracted using the isotropic reorientational eigenmode dynamics (iRED) approach.⁴⁹ In this approach, the following covariance matrix

$$M_{ij} = \frac{1}{2} \langle 3(\vec{\mu}_i \vec{\mu}_j)^2 - 1 \rangle \quad (3)$$

of $\mu(\tau)$ for different N–H vectors was calculated using the Amber 10 ptraj module. The eigenvalues, λ_m , and eigenvectors \vec{m} were then obtained by diagonalization, and the order parameters for residue i were calculated from

$$S_i^2 = 1 - \sum_{m=6}^n \lambda_m |m_i|^2 \quad (4)$$

where the sum runs over all internal modes, i.e., all except those with the five largest eigenvalues, and m_i is the i th element of \vec{m} . Order parameters were calculated either by using the entire trajectory or by averaging over 1 or 5 ns windows. The latter was tested because, if the length of a simulation exceeds the overall tumbling correlation time of the protein, S^2 parameters computed over the whole trajectory can include motions that would not be reflected in the experimental S^2 values, leading to a bias in the computed S^2 values.^{16,50}

Dihedral Distributions. To identify the distribution of dihedral angles of interest, we employed a Gaussian-mixture model (GMM).⁵¹ This approach models the total distribution as a sum of Gaussian (normal) distributions. Each of these distributions will be referred to as a state. In this study, we are only interested in one-dimensional distributions, and hence we employ univariate Gaussian distributions. The probability that a data point (a dihedral angle, denoted y in the following formulas) comes from state k is denoted π_k , and the distribution of each class is

$$p(y|\text{from class } k, \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y - \mu_k)^2}{2\sigma_k^2}\right) \quad (5)$$

where μ_k and σ_k^2 are the mean and variance of state k . The total distribution is

$$p(y|\pi, \mu, \sigma) = \sum_k \pi_k p(y|\text{from class } k, \mu_k, \sigma_k) \quad (6)$$

To determine which state each data point belongs to and the values of the parameters π_k , μ_k , and σ_k^2 , we use an expectation-maximization algorithm.⁵² This is an iterative algorithm that starts with an initial guess of the parameters and then iteratively updates the parameters until convergence. Initially, we assume that there are four states that are uniformly distributed between -180 and $+180$, and that all states have equal probability. If the probability of a state in any iteration falls below 0.001, that state is discarded. In each iteration, the parameters are updated as follows:

$$\begin{aligned}\pi_k &= \frac{1}{n} \sum_{i=1}^n w_{ik} \\ \mu_k &= \frac{1}{n\pi_k} \sum_{i=1}^n w_{ik} y_i \\ \sigma_k^2 &= \frac{1}{n\pi_k} \sum_{i=1}^n w_{ik} (y_i - \mu_k)^2\end{aligned}\quad (7)$$

where the weight w_{ik} is determined from the data and the values of the parameters in the previous iteration (old)

$$w_{ik} = \frac{p(y_i \text{ from class } k, \pi_k^{\text{old}}, \mu_k^{\text{old}}, \sigma_k^{\text{old}}) \pi_k^{\text{old}}}{\sum_k p(y_i \text{ from class } k, \pi_k^{\text{old}}, \mu_k^{\text{old}}, \sigma_k^{\text{old}}) \pi_k^{\text{old}}}\quad (8)$$

Statistical Analysis. Twelve quality measures were employed to judge how well the calculated order parameters (S_{MD}^2) reproduce the measured ones (S_{NMR}^2), viz., the median, the correlation coefficient (r^2), the root mean squared deviation (RMSD), the mean signed deviation (MSD), the mean absolute deviation (MAD), the mean absolute deviation with systematic error removed (MADtr, i.e., after subtraction of the MSD), the mean quote (MQ; $S_{\text{MD}}^2/S_{\text{NMR}}^2$), and the Q value ($Q = (\sum_i (S_{i,\text{MD}}^2 - S_{i,\text{NMR}}^2)^2) / (\sum_i (S_{i,\text{NMR}}^2)^2)$).⁵³ We also calculated how many of the experimental order parameters fall outside the range of the calculated order parameters among the set of simulations of the same type. This measure was also calculated when the range was extended by 0.01, 0.05, and 0.1 in each direction.

Errors in the various qualities were estimated from the standard deviations in S_{MD}^2 and S_{NMR}^2 by performing a random simulation: S_{MD}^2 and S_{NMR}^2 for each residue was assigned a random number from a normal distribution, with the mean and standard deviation obtained in the MD simulations or NMR measurements. Then, we calculated all the quality measures and repeated this procedure 10 000 times. The standard deviations within these sets are reported as the standard error of the quality estimates.

For the comparison of various methods or simulation protocols, we estimate the significance of each prediction by calculating the probability that a certain method will be best the observed number of times or more, using a binomial distribution, assuming equal probability for all methods or simulations. In this calculation, quality measures that give the same results for all methods were omitted.

NMR Relaxation Data. The acquisition and analysis of the NMR relaxation data for the backbone N–H groups have been described.³⁸ In comparing order parameters from NMR and MD, it should be kept in mind that the former depends on assumptions regarding the N–H bond length and chemical shift anisotropy of the ¹⁵N nucleus.¹³ Residue-specific variations in these parameters are not captured by the present approach. Furthermore, for the purposes of the present comparisons, we have also considered the potential effects of additional systematic errors, as follows.

Accurate interpretation of relaxation rates in terms of order parameters requires high-resolution structural information if the protein exhibits anisotropic global rotational diffusion, because the relaxation rates depend on the orientation of the

N–H bond vector in the molecular frame. In the case of Gal3-*apo*, the loops surrounding the saccharide-binding site have different conformations in the low-resolution NMR structure⁵⁴ and the high-resolution X-ray structure, which can be attributed to intermolecular contacts in the crystal.²⁷ In principle, this discrepancy suggests that the experimental S^2 values determined for the loop residues in question might suffer from systematic errors. However, Gal3-*apo* has a modest anisotropy of 1.07, indicating that the potential errors in S^2 should be less than 3%.

The presence of conformational exchange contributions to R_2 requires that the model-free optimization includes an exchange term, R_{ex} . Deviation of the fitted R_{ex} from the actual exchange contribution leads to inaccuracy of the fitted S^2 values. To account for this, in some cases, we have omitted those residues that have been fitted with R_{ex} terms. However, in the case of Gal3, the model-free optimizations appear to be robust. Using reduced data sets excluding R_2 , we obtain nearly identical order parameters for both Gal3-*apo* and Gal3-*lac*, with a weighted RMSD versus the full data sets (including R_2) of 0.007 in both cases.

Result and Discussion

Method to Calculate Order Parameters. Before studying the starting-condition dependence, we addressed which method to use to calculate order parameters. As described in the Methods section, we tested both the ACF and iRED approaches. In the latter case, order parameters were calculated either by using the entire trajectory or by averaging over 1 or 5 ns windows (these methods will be called iRED-full, iRED-1, and iRED-5 in the following). To compare the four methods, we used the 12 quality measures described in the Methods section to judge how well the calculated order parameters (S_{MD}^2) reproduce the measured ones (S_{NMR}^2), taken from our previous investigations of Gal3.³⁸ These comparisons were done both for Gal3-*apo* and Gal3-*Lac*. We also studied the difference in order parameters between Gal3-*apo* and Gal3-*Lac*, ΔS^2 . All the results are collected in Tables S3–S5, in the Supporting Information.

Unfortunately, the various quality measures give different results, as do the simulations on different proteins. The correlation coefficient is in general highest with the iRED-5 method, but the correlation is rather poor for all methods, up to 0.35 and 0.43 for Gal3-*apo* and Gal3-*Lac*, respectively, and less than 0.07 for ΔS^2 . Such a correlation is worse than what has been observed in most previous studies, in which S_{NMR}^2 and S_{MD}^2 have been compared, 0.22–0.93.^{17,18,23,55–60} The reason for this is that r^2 for Gal3 strongly depends on a few residues with a low S^2 , which often are poorly determined by NMR (as will be discussed more below). On the other hand, the RMSD, median, and MAD are actually better than observed in the great majority of previous studies: The RMSD is 0.04–0.06, compared to 0.02–0.26, in the previous studies, with an average of 0.09. In fact, only one investigation in our survey gave an RMSD lower than in the present comparison, 0.02–0.04.⁵⁸ Likewise, both the median (–0.02 to 0.03) and the MAD (0.03–0.04) are lower than in previous studies (0.04 and 0.06–0.11,^{55,56} respec-

Table 1. Residues for which the Four Methods to Calculate S_{MD}^2 Give a Range Larger than 0.05 in the Different Simulations

simulation residue	32M×20 ^a		10A×40 ^a		10A×20 ^a		1A×40 ^a		1A×20 ^a		32M×20 ^b		10M×10 ^b	
	Lac	Apo	Lac	Apo	Lac	Apo	Lac	Apo	Lac	Apo	Lac	Apo	Lac	Apo
Ile115	0.13	0.12	0.18	0.16	0.13	0.13	0.26	0.11	0.16	0.13	0.13	0.12	0.16	0.25
Val116	0.12	0.07	0.12	0.11	0.11	0.09	0.12	0.09	0.12	0.06	0.12	0.07	0.15	0.15
Gly125	0.06	0.06	0.07		0.06		0.09	0.10	0.09	0.12			0.14	0.24
Val126			0.07	0.06					0.07					
Ala142														0.09
Asp154									0.12					
Val155	0.06	0.10	0.10	0.12	0.06	0.08	0.09	0.12		0.07	0.06	0.10	0.11	0.15
Arg168								0.06						
Arg169					0.06				0.07	0.08				
Leu177	0.09	0.08	0.10	0.16	0.09	0.14		0.13		0.23	0.09	0.08	0.18	0.16
Asn179										0.08				
Asn180										0.06				
Arg183														0.08
Glu184				0.06					0.12					0.06
Arg186										0.08				
Val189								0.11		0.11			0.06	0.08
Phe192						0.06			0.06				0.08	0.07
Asp207										0.07				
Val213										0.08				
Ala216													0.06	
Leu219										0.08				
Arg224										0.07	0.06			
Lys227			0.07	0.08						0.07			0.08	0.07
Leu228		0.06	0.07	0.06	0.00	0.06		0.13	0.09	0.06			0.06	0.06
Ile231									0.09					
Ser232	0.10	0.14	0.16	0.15	0.11	0.12	0.24	0.15	0.17	0.13	0.09	0.14	0.18	0.18
Ser246										0.07				
Ile250							0.10	0.10	0.09	0.18				0.07

^a Results based on all four methods. ^b Results based only on the iRED-1 and iRED-full methods.

tively). Thus, the accuracy of the present investigation seems to be similar or better than in previous studies.

The number of NMR values outside the simulated range is typically lowest for the ACF method, but this criterion may favor methods with a poor precision. For the other quality measures, the iRED-1 method gives the best results, at least for the Gal3-apo simulation and the difference. For RMSD, which gives a high weight to outliers, the iRED-full method works better for Gal3-Lac, and for the median, MSD, and MQ, which give a low weight to outliers, ACF performs better for Gal3-Lac. For MAD and MADtr, which give an intermediate weight to outliers, iRED-1 is always best. On the basis of these results, it is hard to point out a single method as the best. In the following, we will use iRED with 1 ns windows, simply because it had the best average performance.

Most importantly, the four different methods give closely similar results for most of the order parameters. In fact, only for seven residues (out of 127), the largest difference among the four methods is larger than 0.05 in the 32 simulation using different starting conformations (32M×20). These residues are listed in Table 1. This shows that, for the great majority of the residues, it does not matter what method is used, whereas for a few residues, different methods give differing results, indicating problems to accurately determine S_{MD}^2 . Thus, a large difference between the four methods can be used as a criterion to decide what residues have a poorly determined S_{MD}^2 , and these could then be excluded from comparisons. As the order parameters calculated with ACF deviate most from those calculated with iRED-1, it is sufficient to calculate order parameters with these two

methods to decide which residues have a poorly determined S_{MD}^2 . However, if the simulation time is short, many of the ACFs will not be converged. Therefore, we recommend using iRED without windowing as a second opinion. From Table 1 (sixth set of columns), it can be seen that this only slightly changes the results.

Conformational Sampling. Next, we turn to simulations started at different structures, based on the alternative conformations in the crystal structure. As detailed in the Methods section and described in Tables S1 and S2 (Supporting Information), we have run 32 simulations of 20 ns length for both Gal3-apo and Gal3-Lac, based on a permutation of five groups of alternative configurations observed in the crystal structure.

A natural question is whether the protein stays in the same conformation during the simulations or if it moves between the various conformations freely. To answer this question, we defined a set of 30 dihedral angles that describe the differences of most of the alternative conformations observed in the crystal structure. They are shown in Table 2.

These dihedral angles were followed throughout the MD simulations. We used a Gaussian-mixture model (GMM)⁵¹ to identify the number of maxima in the distribution function, the dihedral angle at the maxima, and the percent of the time the system spent in each conformation. A typical example is shown in Figure 2.

All 30 dihedral angles describe rotations around a C–C single bond. Therefore, three distinct conformations are expected, rather than the two conformations modeled into the crystal structures. This is confirmed by the simulations: 21 of the 30 angles showed three conformations with a significant probability (>1%; the conformational states

Table 2. Definition of Dihedral Angles Used to Characterize the Conformational Sampling^a

residue	dihedral angle	structure	Apo		Lac		S1	S2	S3
			A	B	A	B			
Val 116	N-CA-CB-CG2	Apo	-89	90	-57		-61	60	
Pro 117	N-CA-CB-CG	Apo	-6	20	23		-23	27	
Asn 119	N-CA-CB-CG	Lac	178		-64	178	-171	-71	77
Arg 129	CA-CB-CG-CD	Lac	-167		175	-153	-180	13	53
Met 130	N-CA-CB-CG	Apo	-160	175	-172		-174	-88	83
Ile 134	CA-CB-CG1-CD1	both	96	172	-55	94	-62	89	170
Lys 139	CB-CG-CD-CE	Apo	-177	71	171		-67	68	179
Pro 140	N-CA-CB-CG	Apo	-28	25	-18		-21	25	
Asn 143	CA-CB-CG-ND2	both	-106	138	-121	132	-161	85	
Gln 150	CA-CB-CG-CD	both	-95	146	-154	153	173		
Arg 168	CB-CG-CD-NE	both	-168	176	-180	167	-69	69	180
Arg 169	CG-CD-NE-CZ	Lac	164		170	-86	-167	-87	148
Val 170	N-CA-CB-CG2	Apo	-175	69	-171		-168	-56	59
Ile 171	N-CA-CB-CG1	Apo	-44	-76	-59		-57	68	
Lys 176	CB-CG-CD-CE	Lac	-164		69	-162	-176	-64	70
Gln 187	CB-CG-CD-NE2	Apo	-122	80	111		-88	96	182
Ser 188	C-CA-CB-OG	both	-70	138	-69	152	-69	174	
Glu 193	CA-CB-CG-CD	Apo	-69	171	-175		-180	-65	59
Lys 196	CA-CB-CG-CD	Lac	-161		-152	171	-175	-64	69
His 208	CA-CB-CG-ND1	Lac	-100		-99	82	-110	97	
Lys 210	CB-CG-CD-CE	Apo	71	-106	76		-66	66	178
Gln 220	C-CA-CB-CG	both	173	-72	172	-61	-62	65	170
Asn 222	CA-CB-CG-ND2	Lac	-71		66	172	-79	72	189
Lys 226	CA-CB-CG-CD	Apo	-64	-179	178		-179	-61	66
Lys 227	CA-CB-CG-CD	Apo	-167	163	166		-180	-67	70
Asn 229	CA-CB-CG-ND2	Apo	-172	168	179		-90	88	162
Ser 232	C-CA-CB-OG	Lac	173		-67	171	-69	179	
Lys 233	C-CA-CB-CG	Apo	165	-163	-145		-61	62	166
Ile 236	CA-CB-CG1-CD1	Apo	176	118	145		-72	91	175
Met 249	C-CA-CB-CG	both	-62	110	-60	-160	-69	71	177

^a The table shows the residue and the atoms in each dihedral angle, the dihedral angles observed in the two conformations in the two crystal structures, as well as the (up to) three conformations observed in the MD simulations, according to the Gaussian-mixture model. Conformations in the crystal structures that belong to the same MD states are marked in bold face. All dihedral angles are in degrees.

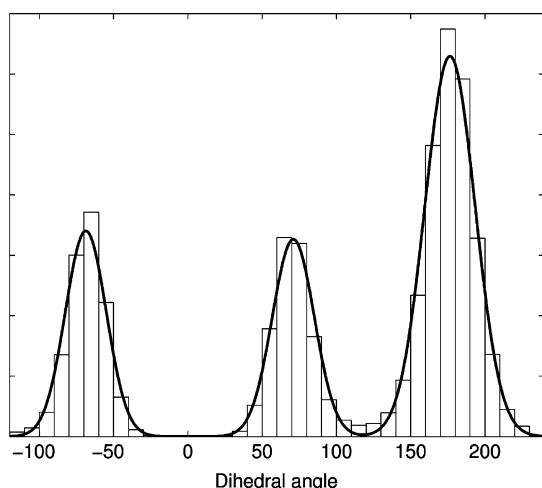


Figure 2. A typical example of the Gaussian-mixture model (GMM) fit for the dihedral angle in Met249. The underlying distribution is shown as a histogram with a bin size of 72 and the GMM is shown as a solid line.

identified with the GMM will be called S1–S3 in the following, whereas the conformations observed in the crystal structures are called A and B). Only for one residue, Gln150, did we find a single significant state (Figure 1 shows the location of important residues in the crystal structure of Gal3). All conformations identified by the GMM are listed in Table 2. It can be seen that most of the conformations identified in the MD simulations are separated by $\sim 120^\circ$,

as expected for a 3-fold rotation. However, nine dihedrals had rotamers that were closer than 100° , down to 41° for S2 and S3 of Arg129.

In general, the dihedral angles with maximum probabilities obtained from the MD simulations are fairly close to those observed in the crystal structure. However, in ~ 5 cases, the difference is over 30° , and for Gln150 in Gal3-apo, crystal conformation A was not observed in the MD simulation. In nine cases, the difference between the two conformations observed in the crystal structures is so small that they belong to the same state in the MD simulations.

If we compare the results of simulations started with all residues in either the A or B conformation, it can be seen that there is a quite large difference between the average values for the 30 dihedral angles in the two simulations, especially for Gal3-apo (Tables S6 and S7, Supporting Information): The averages differ by up to 121° , and the MADs are 30° and 17° for Gal3-apo and Gal3-Lac, respectively. Apparently, the estimated standard errors ($\sim 1^\circ$) for a single simulation grossly overestimate the precision of the averages.

The percentage of the time spent in the various conformations is more stable, but it still shows differences of up to 65% units, with MADs of 11–13 and 8–11% units for Gal3-apo and Gal3-Lac, respectively. This shows that there are significant differences between simulations started from different structures. However, there are also significant dynamics for the studied dihedrals. Only 2–5 dihedrals show a single conformation in the simulations, and for 2–7

additional dihedrals, over 90% of the time is spent in a single conformation. Thus, there is a decent sampling of at least two conformations for most of the dihedrals.

Next, we consider the 32 simulations with the permutations of different conformations. Again, the results (Tables S8 and S9, Supporting Information) show that there is an extensive variation in the results obtained with different starting structures: There are simulations that give completely opposite results for the percentage of the various conformations in the simulations (i.e., some simulations give 100% S1 and others give 100% S2 or S3). This shows that it is mandatory to run several independent simulations to obtain reliable results (or use simulation times much longer than 20 ns).

Interestingly, there is little correlation between what conformations are observed in the crystal structures and the populations of the dihedrals sampled in the simulations. Only in three cases does a residue that has a single conformation in the crystal also populate primarily (>90%) the same conformation in the simulations (Asn119 and His208 in Gal3-apo and Ile171 in Gal3-Lac). For another three dihedrals, MD samples only a single conformation, which essentially covers the two states observed in the crystal structures (i.e., the two conformations observed in the crystal structures are so close that they belong to the same MD conformation). On the other hand, there are three residues that show almost only one conformation in the simulations, but two conformations in the crystal structures (Val116 in Gal3-apo and Asn119 and Asn222 in Gal3-Lac). All of the other residues show two or three conformations in the MD simulations, independently of the number of conformations observed in the crystal structures. The reason for this may be that the resolution of the structures is too low to discern several conformations (some of which have a low occupancy), that crystal packing effects may stabilize certain conformations, or that the low temperature conditions during the X-ray diffraction experiments (employing liquid nitrogen) restricts the number of populated conformational states.

There are some conspicuous differences between Gal3-apo and Gal3-Lac. In particular, Asn222 is almost entirely in the S3 state in the Lac simulation, whereas it is only 1% of the time in that conformation for Gal3-apo. Glu193 and His208 also show quite large differences. However, for most of the dihedrals, the occupancy of the various conformations is quite similar, with a MAD of only 8% units.

The prime question in this investigation is whether it is necessary to use different starting structures to obtain a proper sampling or if similar results can be obtained with different means. To this end, we compare the 32 simulations started from different conformations with 10 simulations started from the same structure (all residues in the A conformation), but with different starting velocities. The results in Tables S8 and S9, Supporting Information, show that the 10 simulations started from the same conformation show a slightly smaller sampling. For example, the average range of the percentages (maximum – minimum) of the three states of the dihedrals is 24–32% units for the 10 simulations, but 37–50% units for the 32 simulations. On the other hand, the average standard deviation of the percentage of the three conformations is similar or slightly larger for the 10

simulations, 2–3% units. However, the MAD between the 32 and 10 simulations is only ~4% units for all three states in both proteins, and these numbers are dominated by three residues from each simulation, (Pro140, Glu193, and Asn222 for Gal3-apo and His208, Gln220, and Lys233 for Gal3-Lac), which show differences of 14–30% units. Thus, there is some advantage of starting the simulations from different conformations, but the effect is rather small.

Another interesting question is how long simulations are needed for converged results. In Tables S10 and S11 (Supporting Information), we compare the results obtained for the 10 independent simulations after 20 and 40 ns of simulation time. It can be seen that the results in general are similar, with MADs of 2–3% units for both complexes, and with maximum differences of up to 7–11% units. They also give similar differences, compared to the 32 simulations. Thus, we can conclude that the conformational sampling is reasonably converged already after 20 ns.

Related to this issue is the time-scale of the conformational changes studied. If it is short, compared to the simulation time, the results should be converged, and then it should also be possible to estimate the equilibrium constants from the observed percentage and the activation barriers from the time-scale. In Tables S12 and S13 (Supporting Information), we therefore list how long it takes before the protein changes the conformation of the various dihedrals. It can be seen that the time varies from 4 ps for Ser188 to 2.8 ns for His208, with an average of ~0.7 ns for both systems. This indicates that the sampling of 20–40 ns should be appropriate, although the sampling of states with a low occupancy can be worse. For most residues, the simulations starting from 10 different velocities or 32 different structures give a similar result, but for 3–4 residues, the difference is large, up to 6 ns.

Starting-Condition Dependence of Order Parameters.

Next, we consider order parameters obtained from the 32 + 10 different simulations of Gal3-apo and Gal3-Lac. Our prime question is how to perform simulations that give the best results, compared to experiments. The results of the 12 different quality measures used to compare S_{NMR}^2 to S_{MD}^2 are listed in Table 3 for five different sets of simulations, viz., the 32 simulations with different starting structures (20 ns length; 32M×20), the 10 independent simulations starting from the same structure (of either 40 or 20 ns length; 10A×40 and 10A×20), and a single simulation of either 40 or 20 ns length (1A×40 and 1A×20). Figure 3 shows the ranking of the various simulations, i.e., the number of times each of the simulations rank first, second, and so on, for the various quality measures and systems.

From Figure 3, it is clear that the 32M×20 simulation is best: It gives the best results for all measures, except the median, for both complexes, as well as for the difference. If we take into consideration the uncertainties in the various quality estimates (both from NMR and MD), the 32M×20 simulation gives significantly better results at the 95% level (according to a Student's *t* test) for 2–6 quality measures compared to the other four simulations (Table S14, Supporting Information).

Table 3. Comparison of S_{MD}^2 and S_{NMR}^2 for Gal3-apo, Gal3-Lac, and the Difference between the Two Proteins^a

simulation	RMSD	r^2	MAD	MADtr	MSD	median	MQ	Q	$n \pm 0$	$n \pm 0.01$	$n \pm 0.05$	$n \pm 0.1$
Gal3-apo												
32M×20	0.038	0.28	0.029	0.029	-0.004	0.003	1.00	0.002	63	46	6	0
10A×40	0.039	0.24	0.030	0.031	-0.004	0.000	1.00	0.002	80	61	9	1
10A×20	0.039	0.26	0.030	0.030	-0.004	-0.001	1.00	0.002	79	57	8	0
1A×40	0.041	0.21	0.031	0.031	-0.004	0.000	1.00	0.002				
1A×20	0.044	0.20	0.033	0.033	-0.004	0.003	1.00	0.003				
Gal3-Lac												
32M×20	0.062	0.35	0.041	0.036	0.028	0.026	1.04	0.005	76	59	15	2
10A×40	0.063	0.33	0.041	0.037	0.028	0.026	1.04	0.005	88	73	23	4
10A×20	0.063	0.33	0.041	0.037	0.028	0.025	1.04	0.006	83	70	20	4
1A×40	0.065	0.29	0.042	0.037	0.029	0.026	1.04	0.006				
1A×20	0.066	0.28	0.043	0.038	0.030	0.025	1.04	0.006				
difference												
32M×20	0.054	0.05	0.036	0.027	0.032	0.025						
10A×40	0.057	0.01	0.036	0.028	0.032	0.026						
10A×20	0.057	0.01	0.036	0.028	0.032	0.026						
1A×40	0.059	0.02	0.038	0.030	0.032	0.024						
1A×20	0.062	0.02	0.040	0.031	0.034	0.026						

^a The 12 different quality measures listed are the root-mean-squared-deviation (RMSD), Pearson's correlation coefficient (r^2), the mean absolute deviation (MAD), the mean absolute deviation when removing the systematic error (MADtr), the mean signed deviation (MSD), the median, the mean quote (MQ; S_{MD}^2/S_{NMR}^2), the Q value, and the number of residues for which the S_{NMR}^2 value falls outside the range of the S_{MD}^2 values (when there are several simulations; $n \pm 0$). The latter measure is also calculated when the MD range is extended by 0.01, 0.05, and 0.1 in each direction ($n \pm 0.01$, $n \pm 0.05$, and $n \pm 0.1$). The best result for each quality measure for each system is marked in bold face. The iRED method with 1 ns windows was used to obtain S_{MD}^2 , and the equilibration time was 5 ns.

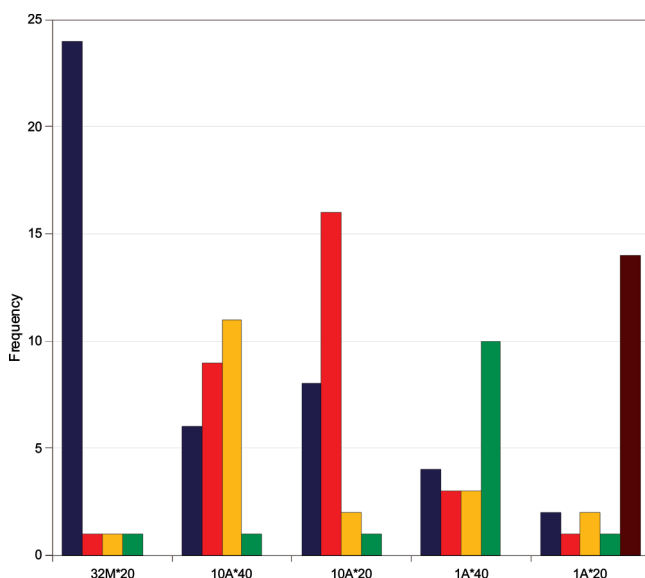


Figure 3. Ranking of the five simulations for the 12 quality measures and three systems (Gal3-apo, Gal3-Lac, and the difference between the two systems). The figure shows the number of times each simulation ranks as number one, two, three, four, or five (from left to right) for each quality measure and simulation.

However, it is also notable how small the differences are: If we instead use a single simulation of 20 ns length, the RMSD increases by up to 0.006, the MAD by 0.004, the MSD by 0.002, and Q by 0.001, whereas the median and MQ hardly change. It is only the correlation coefficient that increases by a larger amount, up to 0.08. Compared to the 10 independent simulations, the differences are even smaller, and the main difference is seen for the total range of the simulated values; that is, it is appreciably larger for the 32 different starting structures (illustrated by a decreased number of S_{NMR}^2 values outside the S_{MD}^2 range).

Interestingly, there seems to be little gain in running the 10 independent simulations for a longer time (40 ns, rather than 20 ns): Only two quality measures are improved and two become worse, in all cases by a minimal amount. Moreover, the number of S_{NMR}^2 values outside the range of the calculated S_{MD}^2 range increases, although this only illustrates that it is a poor quality measure, favoring simulations with a poor precision. However, for the single simulation, most of the quality measures are improved if the simulation is prolonged from 20 to 40 ns. Thus, we can conclude that there is small, but consistent, improvement in the results as more simulations are performed. It seems more favorable to run several shorter simulations than one long one.

The analysis above is based on sets of simulations that have different total simulation times. To make a more fair comparison, we devised new sets of simulations, which have a total simulation time of either 20 or 40 ns. Until now, we have used an equilibration time of 5 ns, a decision that was made on the basis of a rather qualitative analysis of the backbone RMSD fluctuations. We therefore studied the effect of equilibration times that ranged from 0 to 19 ns for the 32 simulations of mixed conformations. The results in Table 4 show that the $S_{NMR}^2 - S_{MD}^2$ differences are rather insensitive to the equilibration time: Only the correlation coefficient (and RMSD for $\Delta\Delta S^2$) show significant differences between 0 and 19 ns equilibration times. This is because the correlation coefficient is very sensitive to the actual S^2 value of a few vectors (all except 1–4 order parameters are between 0.7 and 1.0 for both NMR and MD). For example, r^2 for $\Delta\Delta S^2$ can increase from 0.04 to 0.24 upon a change of only 0.08 for a single residue in one of the simulations. On the basis of these results, we decided to use an equilibration time of 0.25 ns.

Next, we created seven new sets of simulations, all using an equilibration time of 0.25 ns: Three of them have a total

Table 4. Equilibration-Time Dependence of the Quality Measures for Gal3-apo, Gal3-Lac, and the Difference, Compared to NMR^a

time (ns)	Gal3-apo										Gal3-Lac										Difference			
	RMSD	r ²	MAD	MADTr	MSD	Med	MQ	Q	RMSD	r ²	MAD	MADTr	MSD	Med	MQ	Q	RMSD	r ²	MAD	MADTr	MSD	Med		
0	0.038	0.409	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.358	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.040	0.036	0.027	0.032	0.025		
0.1	0.038	0.409	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.357	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.037	0.036	0.027	0.032	0.025		
0.2	0.038	0.409	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.357	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.035	0.036	0.027	0.032	0.025		
0.3	0.038	0.408	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.357	0.041	0.036	0.028	0.027	1.040	0.005	0.054	0.037	0.036	0.027	0.032	0.025		
0.4	0.038	0.408	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.357	0.041	0.036	0.028	0.027	1.040	0.005	0.054	0.037	0.036	0.027	0.032	0.025		
0.5	0.038	0.408	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.357	0.041	0.036	0.028	0.027	1.040	0.005	0.054	0.038	0.036	0.027	0.032	0.025		
0.7	0.038	0.409	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.358	0.041	0.036	0.028	0.027	1.040	0.005	0.054	0.047	0.036	0.027	0.031	0.025		
0.9	0.038	0.409	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.358	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.040	0.036	0.027	0.032	0.025		
1	0.038	0.410	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.357	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.046	0.036	0.027	0.032	0.024		
2	0.038	0.410	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.356	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.050	0.036	0.027	0.032	0.024		
3	0.038	0.409	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.355	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.057	0.036	0.027	0.032	0.024		
4	0.038	0.408	0.029	0.029	-0.003	0.002	1.000	0.002	0.062	0.353	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.057	0.036	0.027	0.032	0.024		
5	0.038	0.408	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.351	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.057	0.036	0.027	0.032	0.024		
6	0.038	0.409	0.029	0.029	-0.004	0.002	1.000	0.002	0.062	0.348	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.052	0.036	0.027	0.032	0.024		
7	0.038	0.408	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.345	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.044	0.036	0.027	0.032	0.024		
8	0.038	0.408	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.343	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.044	0.036	0.027	0.032	0.024		
9	0.038	0.407	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.340	0.041	0.037	0.028	0.026	1.040	0.005	0.054	0.045	0.036	0.027	0.032	0.023		
10	0.038	0.403	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.339	0.041	0.037	0.028	0.026	1.040	0.005	0.054	0.050	0.036	0.027	0.032	0.023		
11	0.038	0.405	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.338	0.041	0.037	0.028	0.026	1.040	0.005	0.054	0.050	0.036	0.027	0.031	0.023		
12	0.038	0.405	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.344	0.041	0.036	0.028	0.026	1.040	0.005	0.054	0.047	0.036	0.027	0.032	0.023		
13	0.038	0.401	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.348	0.041	0.036	0.028	0.025	1.040	0.005	0.054	0.060	0.036	0.027	0.032	0.024		
14	0.038	0.402	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.349	0.041	0.036	0.028	0.025	1.040	0.005	0.054	0.082	0.036	0.027	0.032	0.023		
15	0.038	0.407	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.349	0.041	0.036	0.028	0.025	1.040	0.005	0.054	0.084	0.036	0.027	0.032	0.023		
16	0.038	0.400	0.029	0.029	-0.004	0.002	1.000	0.002	0.062	0.349	0.041	0.036	0.028	0.026	1.040	0.005	0.053	0.108	0.036	0.027	0.031	0.023		
17	0.038	0.395	0.029	0.029	-0.004	0.003	1.000	0.002	0.062	0.351	0.041	0.036	0.028	0.026	1.040	0.005	0.053	0.122	0.036	0.027	0.031	0.024		
18	0.039	0.374	0.029	0.030	-0.003	0.002	1.000	0.002	0.062	0.354	0.041	0.036	0.028	0.026	1.040	0.005	0.052	0.172	0.035	0.026	0.031	0.023		
19	0.039	0.370	0.030	0.030	-0.003	0.003	1.000	0.002	0.061	0.381	0.040	0.036	0.028	0.026	1.040	0.005	0.051	0.245	0.036	0.026	0.032	0.024		

^a The quality measures are the same as in Table 3.

Table 5. Comparison of S_{MD}^2 and S_{NMR}^2 for Gal3-apo, Gal3-Lac, and the Difference between the Two Proteins^a

set	Gal3-apo										Gal3-Lac										Difference			
	RMSD	r^2	MAD	MADtr	MSD	Med	MQ	Q	RMSD	r^2	MAD	MADtr	MSD	Med	MQ	Q	RMSD	r^2	MAD	MADtr	MSD	Med		
1A×20	0.041	0.23	0.031	0.031	-0.003	0.003	1.00	0.002	0.066	0.28	0.043	0.038	0.030	0.027	1.04	0.006	0.062	0.02	0.040	0.032	0.033	0.027		
10A×2	0.041	0.18	0.031	0.032	-0.003	0.004	1.00	0.002	0.065	0.30	0.043	0.037	0.031	0.026	1.04	0.006	0.061	0.08	0.038	0.029	0.034	0.027		
10M×2	0.039	0.20	0.030	0.031	- 0.002	0.004	1.00	0.002	0.063	0.36	0.042	0.035	0.031	0.028	1.04	0.006	0.057	0.02	0.038	0.028	0.033	0.026		
1A×40	0.039	0.22	0.030	0.030	-0.003	0.001	1.00	0.002	0.065	0.29	0.042	0.037	0.029	0.026	1.04	0.006	0.059	0.03	0.038	0.030	0.032	0.023		
10A×4	0.040	0.21	0.031	0.031	-0.003	0.002	1.00	0.002	0.064	0.33	0.042	0.036	0.030	0.025	1.04	0.006	0.061	0.11	0.037	0.029	0.033	0.025		
32M×1	0.039	0.19	0.030	0.031	-0.003	0.001	1.00	0.002	0.063	0.36	0.041	0.035	0.031	0.027	1.04	0.005	0.058	0.00	0.038	0.028	0.033	0.025		
10M×4	0.038	0.23	0.030	0.030	-0.003	0.003	1.00	0.002	0.062	0.37	0.041	0.035	0.030	0.027	1.04	0.005	0.057	0.02	0.037	0.028	0.033	0.026		
10A×5	0.039	0.38	0.030	0.031	-0.003	0.002	1.00	0.002	0.064	0.33	0.042	0.036	0.030	0.025	1.04	0.006	0.061	0.16	0.037	0.029	0.033	0.023		
5A×10	0.039	0.37	0.030	0.030	-0.003	0.004	1.00	0.002	0.063	0.34	0.041	0.036	0.030	0.025	1.04	0.006	0.060	0.08	0.037	0.029	0.032	0.025		
10M×5	0.038	0.40	0.029	0.030	-0.003	0.003	1.00	0.002	0.062	0.37	0.041	0.035	0.030	0.026	1.04	0.005	0.056	0.02	0.037	0.027	0.033	0.026		
5M×10	0.038	0.40	0.029	0.030	-0.003	0.002	1.00	0.002	0.062	0.37	0.041	0.036	0.029	0.026	1.04	0.005	0.055	0.05	0.037	0.028	0.032	0.025		
10A×10	0.039	0.38	0.030	0.030	- 0.003	0.003	1.00	0.002	0.063	0.34	0.041	0.036	0.029	0.025	1.04	0.005	0.059	0.10	0.036	0.029	0.032	0.023		
5A×20	0.039	0.38	0.030	0.030	-0.004	0.001	1.00	0.002	0.063	0.33	0.041	0.037	0.029	0.025	1.04	0.006	0.058	0.03	0.037	0.028	0.032	0.025		
10M×10	0.038	0.41	0.029	0.029	- 0.003	0.002	1.00	0.002	0.061	0.37	0.041	0.036	0.029	0.026	1.04	0.005	0.055	0.04	0.036	0.027	0.032	0.025		
5M×20	0.038	0.40	0.029	0.030	- 0.003	0.002	1.00	0.002	0.062	0.35	0.041	0.036	0.028	0.026	1.04	0.005	0.055	0.04	0.036	0.028	0.032	0.025		
32M×5	0.038	0.40	0.029	0.029	-0.003	0.003	1.00	0.002	0.062	0.37	0.041	0.035	0.030	0.025	1.04	0.005	0.056	0.00	0.037	0.027	0.032	0.025		
16M×10	0.038	0.41	0.029	0.029	-0.003	0.002	1.00	0.002	0.061	0.37	0.041	0.036	0.029	0.026	1.04	0.005	0.055	0.03	0.036	0.027	0.032	0.025		
32M×10	0.037	0.41	0.029	0.029	-0.003	0.003	1.00	0.002	0.061	0.37	0.041	0.036	0.029	0.026	1.04	0.005	0.055	0.03	0.036	0.027	0.032	0.025		
16M×20	0.038	0.41	0.029	0.029	-0.003	0.002	1.00	0.002	0.062	0.36	0.041	0.036	0.028	0.026	1.04	0.005	0.054	0.04	0.036	0.027	0.032	0.025		

^a The quality measures are the same as in Table 3. The best results within each group are marked in bold face. The equilibration time was always 0.25 ns.

simulation time of 20 ns. In the first, we take a single 20 ns simulation, started with the all-A conformation (1A×20). In the second, we instead take 10 independent simulations of 2 ns, all started from the A conformations (10A×2). In the third set, we take 10 simulations of 2 ns, started from different conformations (10M×2). These 10 simulations can be selected in many ways from the 32 simulations we have run with different starting conformations. We simply selected 10 simulations out of these 32 at random and repeated this 50 times to obtain a stable average. From the results in Table 5, it can be seen that the third set (10M×2) gives slightly better results than the other two sets: It gives the best result for 16 of the 22 quality criteria examined (we did not consider here the number of residues for which the S_{NMR}^2 value falls outside the range of the S_{MD}^2 values because our previous results indicated that it is a poor quality measure). The probability that we would get such a result if the distribution was completely random is less than 3%. The other two sets were best only for seven or nine quality measures. Thus, it is better to run 20 short simulations than one long one, and it is also better to start from several different conformations than a single one.

Likewise, we constructed four sets of simulations with a total length of 40 ns. The first is a 40-ns simulation started from a single conformation (1A×40). The second is 10 independent simulations of 4 ns, all started from the same conformation (10A×4). The third is 32 simulations of 1.25 ns, started from different conformations, (32M×1), whereas the fourth is 10 simulations of 4 ns length, started from different conformations (10M×4; again an average over 50 different random selections of 10 simulations out of the available 32 different simulations). The results in Table 5 indicate that there is a slight advantage to start with different conformations: The 10M×4 simulations gave the best results for 16 quality measures (89% significance), whereas the second best methods 1A×40 and 32M×1.25 are best for 10 quality measures. The last method, 10A×4 is best for seven quality measures. This also indicates that 1.25 ns is a too short a time for the simulation of order parameters—the 4 ns simulations give better results, even if fewer simulations are run. This result is expected, because the experimentally determined correlation time for the rotational diffusion is 7–8 ns for both proteins.³⁸

Comparing the results with 20 or 40 ns total simulation time, there is a clear improvement when using the longer simulation time for 12 of the quality criteria, and only one becomes worse (>99% significance). Likewise, there is a clear improvement in the results going from the best set of simulations with a total time of 40 ns and the 32M×20 simulations in Table 3: 11 quality criteria are improved, especially for the difference between the two proteins, whereas only three become worse (97% significance). This shows that the order parameters can be improved by extending the simulations, although the convergence is very slow.

This observation led us to continue the investigation with simulations of a total length of 50, 100, 160, and 320 ns. The results are also included in Table 5. It can be seen that, for a total simulation time of 50 ns, it is better to run five 10 ns simulations than 10 5-ns simulations (95% significance),

irrespective of whether they are started from a single or many different conformations, although the latter gives the best results (89% significance). On the other hand, for a total simulation time of 100 ns, it is better to run 10 10-ns simulations than five 20 ns simulations (89–94% significance). Again, simulations started from several conformations give the best results (95% significance). These results are confirmed for the even longer total simulation time: It is better to run 16 simulations of a length of 10 ns than 32 simulations of 5 ns length (95% significance). Only for the longest simulation time (320 ns) do the results become inconclusive—there is no significant difference between 32 simulations of 10 ns length or 16 simulations of 20 ns length. However, the conclusion remains that there is no advantage of running the longer simulations. Therefore, we can with good confidence conclude that the optimum simulation length, at least for Gal3, is ~10 ns.

It can also be seen from Table 5 that we reach convergence for the various quality measures. Between 50 and 100 ns total simulation time, there is an improvement for 10 of the quality measures, whereas only one is worse (for the two best methods; 98% significance). However, going from 100 to 160 or 320 ns total simulation time, there is no longer any clear improvement. In fact, by comparing the results in Tables 3 and 5, it can be seen that the 10M×10 simulations actually give better results than the full 32M×20 simulations for six of the quality measures and only two of them give worse results. This also confirms that we can use a short equilibration time of 0.25 ns. Therefore, we conclude that, for Gal3, the ideal simulation protocol involves 10 simulations of 10 ns, starting from different conformations.

There is also a difference between the various types of simulations with regard to the stability of the calculated S_{MD}^2 , as estimated from the difference in results obtained with the four methods to estimate S_{MD}^2 . From Table 1, it can be seen that, with a single 20 ns simulation, 24 residues have a range larger than 0.05 between the various methods, and most of them are only observed for one protein. This number is decreased to 10 if the simulation is extended to 40 ns, and a similar number is observed also for the 10 independent simulations started from the same structure, irrespective of whether they are 20 or 40 ns long. However, for the 32 simulations started from different conformations, only 6–7 residues have poorly determined S_{MD}^2 's. Thus, starting from several different conformations makes the results more stable and well-determined. However, this also depends on the length of the simulations. If we instead use only 10M×10 simulations (with 0.25 ns equilibration), there are 15 residues with poorly determined S_{MD}^2 's, although this set can be reduced to something similar to the 32M×20 set by using a threshold of ~0.09 instead.

For seven residues, the range of S_{MD}^2 obtained with different methods is large in all simulation sets, viz., Ile115, Val116, Gly125, Val155, Leu177, Leu228 (not in Gal3-Lac), and Ser232 (cf. Figure 1). Only two of these, Val116 and Ser232, have different conformations in the crystal structures, but they both reside mainly in one conformation (S1) during the MD simulations (74–91%).

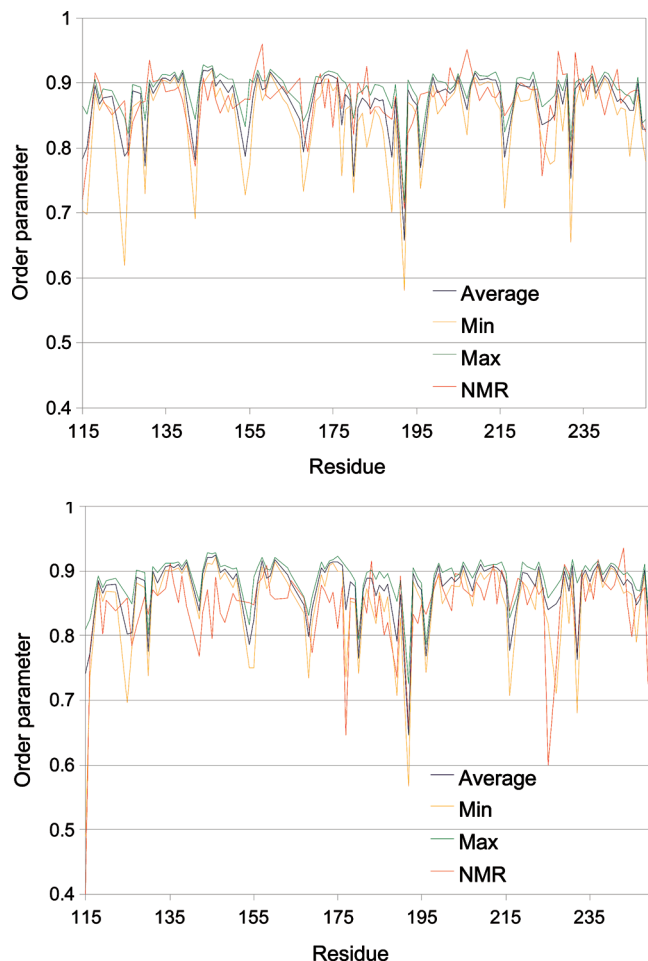


Figure 4. S_{MD}^2 parameters from the $32\text{M} \times 20$ ns simulations (average, minimum, and maximum over the 32 simulations), compared to the corresponding S_{NMR}^2 parameters for (top) Gal3-apo and (bottom) Gal3-Lac. Only residues with S_{NMR}^2 parameters in both states are shown.

In Figure 4, we compare the calculated and measured order parameters for Gal3-apo and Gal3-Lac. The maximum value of each order parameter over the 32 simulations is always close to the average, whereas the minimum shows a rather large variation for some of the residues. The residues that have the largest absolute difference between S_{MD}^2 and S_{NMR}^2 are Lys196, Met130, Asp207, Asp154, and Gly125 for Gal3-apo and Ile115, Val225, Leu177, Ile145, and Met249 for Gal3-Lac. As can be seen in Figure 5, large differences are primarily observed in loops in the protein structure, whereas the β sheets are well described. Three of the residues with large deviations, Ile115, Gly125, and Leu177, have poorly determined S_{MD}^2 's, whereas the other residues with poorly determined S_{MD}^2 's do not show any conspicuous errors. Interestingly, residues with large errors have a too high S_{MD}^2 for Gal3-Lac, but a too low S_{MD}^2 for Gal3-apo. The errors are also larger for Gal3-Lac (up to 0.35) than for Gal3-apo (up to 0.10).

If the residues with poorly determined S_{MD}^2 's are removed from the analysis, the results are significantly improved for Gal3-Lac and for $\Delta\Delta S^2$, as can be seen in Table 6 (97% significance). For example, the MADs decrease from 0.041 and 0.036 to 0.037 and 0.032. On the other hand, correlation coefficients become worse, simply because the poorly

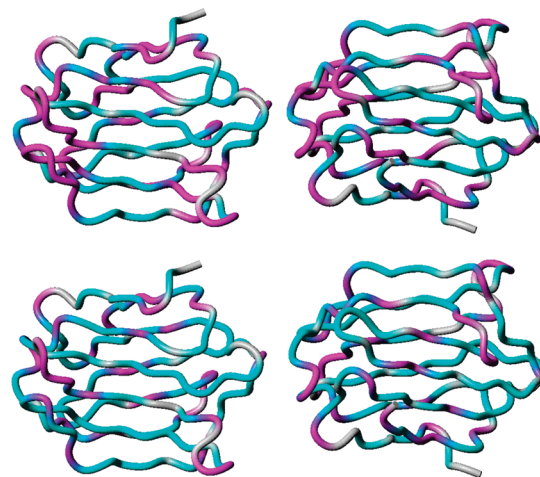


Figure 5. Mapping of the $S_{\text{MD}}^2 - S_{\text{NMR}}^2$ difference onto the crystal structure of Gal3. Two pictures are shown of each Gal3-apo (top) and Gal3-Lac (bottom), related by a 180° rotation. The scale runs from dark blue ($S_{\text{MD}}^2 < S_{\text{NMR}}^2$) via cyan ($S_{\text{MD}}^2 = S_{\text{NMR}}^2$) to magenta. Gray color indicates that data are missing.

determined residues typically have low S^2 's, contributing strongly to the correlation coefficient. Further improvement is seen if residues in loops surrounding the saccharide-binding site are omitted or if residues that have been fitted with a R_{ex} term in the NMR experiments are omitted, as is also seen in Table 6 (for example, r^2 increases to 0.54–0.70), but then, the number of considered residues becomes rather small (41). We have also checked whether residues that have two conformations in the crystal structure, or are located close to such residues, give worse results when S_{MD}^2 is compared to S_{NMR}^2 . However, we did not find any such trends.

Finally, we have also included in the table the results of the simple contact model, suggested by Zhang and Brüschweiler.⁶¹ It can be seen that it gives worse results for all quality measures, except the RMSD, MSD, MQ, and median for Gal3-Lac and r^2 for $\Delta\Delta S^2$ (significance 97%). Thus, the simulations provide significantly improved predictions of the order parameters compared to the contact model.

Conclusions

In this paper, we have addressed a number of questions of importance in the calculation of backbone N–H order parameters from MD simulations. First, we have compared four different methods to extract the S_{MD}^2 parameters, viz., ACF and iRED with three different window sizes. Different quality measures give different results, as do different simulated systems, so we cannot reach any definite conclusions. ACF seems to give the results with the largest spread; i.e., it seems to be more sensitive to the convergence of the simulations than the iRED approach. The iRED method with windows of 1 ns seems to give the best precision and the smallest outliers on average, but the median and correlation coefficient were sometimes worse than for other variants of iRED. However, the most important result was that the four methods gave similar results for most of the studied S_{MD}^2 parameters, indicating that all methods give reliable results. In fact, if the four methods differ significantly (e.g., by more than 0.05), it indicates that

Table 6. Comparison of S_{MD}^2 and S_{NMR}^2 for Gal3-apo, Gal3-Lac, and the Difference between the Two Proteins When Various Residues Are Omitted from the Comparison^a

simulation	<i>n</i>	RMSD	r^2	MAD	MADtr	MSD	median	MQ	Q
Gal3-apo									
32M×20	109 ^b	0.038	0.28	0.029	0.029	−0.004	0.003	1.00	0.002
	104 ^c	0.038	0.34	0.029	0.029	−0.004	0.003	1.00	0.002
	82 ^d	0.029	0.60	0.023	0.023	−0.001	0.004	1.00	0.001
	41 ^b	0.030	0.68	0.023	0.023	−0.007	−0.004	0.99	0.001
10M×10	109 ^b	0.038	0.41	0.029	0.029	−0.003	0.002	1.00	0.002
	104 ^c	0.037	0.34	0.029	0.029	−0.003	0.002	1.00	0.002
	82 ^d	0.029	0.59	0.023	0.023	0.000	0.004	1.00	0.001
	41 ^e	0.029	0.70	0.023	0.023	−0.006	−0.003	0.99	0.001
contact model	109 ^b	0.073	0.17	0.049	0.046	−0.034	−0.018	0.96	0.007
Gal3-Lac									
32M×20	109 ^b	0.062	0.35	0.041	0.036	0.028	0.026	1.04	0.005
	104 ^c	0.050	0.31	0.037	0.032	0.025	0.026	1.03	0.003
	82 ^d	0.052	0.48	0.031	0.028	0.023	0.022	1.03	0.004
	41 ^e	0.066	0.50	0.035	0.034	0.025	0.022	1.04	0.006
10M×10	109 ^b	0.061	0.37	0.041	0.036	0.029	0.026	1.04	0.005
	104 ^c	0.050	0.31	0.037	0.032	0.025	0.026	1.03	0.003
	82 ^d	0.052	0.49	0.032	0.027	0.024	0.022	1.03	0.004
	41 ^e	0.065	0.54	0.035	0.033	0.027	0.022	1.04	0.006
contact model	109 ^b	0.058	0.31	0.041	0.040	0.003	0.005	1.01	0.005
difference									
32M×20	109 ^b	0.054	0.05	0.036	0.027	0.032	0.025		
	104 ^c	0.042	0.00	0.032	0.022	0.028	0.025		
	82 ^d	0.044	0.22	0.029	0.021	0.024	0.019		
	41 ^e	0.057	0.23	0.035	0.025	0.032	0.023		
10M×10	109 ^b	0.055	0.04	0.036	0.027	0.032	0.025		
	104 ^c	0.043	0.01	0.033	0.022	0.029	0.025		
	82 ^d	0.045	0.17	0.029	0.021	0.024	0.021		
	41 ^e	0.058	0.18	0.036	0.025	0.033	0.025		
contact model	109 ^b	0.075	0.09	0.042	0.034	0.037	0.025		

^a *n* is the number of residues included in the comparison. The quality measures are the same as in Table 3. ^b All residues are included. ^c Five poorly determined S_{MD}^2 's according to the 32M×20 simulations were omitted (cf. Table 1). ^d Residues in loops surrounding the saccharide-binding site are omitted. ^e Residues that have been fitted with a R_{ex} term are omitted.

there are problems with the convergence of the calculations; consequently, we suggest that it is good practice to exclude the affected residues from detailed interpretation.

Second, we have studied how the calculated S_{MD}^2 parameters depend on the starting conditions of the MD simulations. It is clearly inappropriate to base the calculations on a single MD simulation. Better results are obtained if the results of several independent simulations are averaged. They can be obtained by simply using different starting velocities, but it is advantageous to use several different conformations, if present in the crystal structure.

Third, we have compared different lengths of the simulations. Our calculations show that, at least for Gal3, the results are better if the simulation length is increased from 5 to 10 ns, but there is no significant improvement if they are extended to 20 or 40 ns (keeping the total simulation time constant by running several independent simulations). Moreover, there is no significant improvement when extending the total simulation time over 100 ns, except that a few order parameters become better determined. Thus, our results indicate that the ideal simulation protocol is 10 independent simulations of 10 ns length, started from different conformations.

Fourth, even if the RMSD of the coordinates indicates that an equilibration time of 5 ns is needed to reach stable results, this has a small influence on the calculated S_{MD}^2 parameters, at least when averaged over several independent simulations. In fact, 10 × 10 ns simulations with an equilibration time

of only 0.25 ns give as good results as 32 × 20 ns simulations with 5 ns equilibration.

Fifth, it should be noted that, even after 400–640 ns simulation time, the correspondence between calculated and measured S^2 parameters is rather poor, with a correlation coefficient of less than 0.43, a MAD of over 0.029, and with a maximum error of up to 0.35.

Finally, although this study has concentrated on a comparison of calculated and measured S^2 order parameters, we are confident that most of our conclusions are applicable also to calculations of other properties from MD simulations, as other investigations indicate.²⁵

Supporting Information Available: Description of the selection of alternative conformations; description of the alternative conformations and their distinct hydrogen-bond patterns; comparison of the four methods to obtain S_{MD}^2 ; description of the dihedral conformations observed in the various simulations; transition times between the various dihedral conformations; and comparison of the various simulations in Table 3, taking into account the statistical uncertainties in both S_{MD}^2 and S_{NMR}^2 . This information is available free of charge via the Internet at <http://pubs.acs.org/>.

Acknowledgment. This investigation has been supported by grants from the Swedish research council and from the Research School in Pharmaceutical Science. It has also

been supported by computer resources of Lunarc at Lund University and at HPC2N at Umeå University.

References

- (1) Akke, M.; Brüschweiler, R.; Palmer, A. G. NMR order parameters and free energy: An analytical approach and its application to cooperative Ca²⁺ binding by calbindin D_{9k}. *J. Am. Chem. Soc.* **2003**, *115*, 9832–9833.
- (2) Homans, S. W. Probing the binding entropy of ligand-protein interactions by NMR. *ChemBioChem* **2005**, *6*, 1585–1591.
- (3) Igumenova, T. I.; Frederick, K. K.; Wand, A. J. Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. *Chem. Rev.* **2006**, *106*, 1672–1699.
- (4) Lipari, G.; Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.* **1982**, *104*, 4546–4559.
- (5) Clore, G. M.; Szabo, A.; Bax, A.; Kay, L. E.; Driscoll, P. C.; Gronenborn, A. M. Deviations from the simple two-parameter model-free approach to the interpretation of ¹⁵N nuclear magnetic relaxation of proteins. *J. Am. Chem. Soc.* **1990**, *112*, 4989–4991.
- (6) Halle, B. The physical basis of model-free analysis of NMR relaxation data from proteins and complex fluids. *J. Chem. Phys.* **2009**, *131*, 224507.
- (7) Palmer, A. G. NMR probes of molecular dynamics: overview and comparison with other techniques. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 129–155.
- (8) Boyd, J. Measurement of ¹⁵N Relaxation Data from the Side Chains of Asparagine and Glutamine Residues in Proteins. *J. Magn. Reson. B* **1995**, *107*, 279–285.
- (9) Berglund, H.; Baumann, H.; Knapp, S.; Ladenstein, R.; Härd, T. Flexibility of an Arginine Side Chain at a DNA-Protein Interface. *J. Am. Chem. Soc.* **1995**, *117*, 12883–12884.
- (10) Muhandiram, D. R.; Yamazaki, T.; Sykes, B. D.; Kay, L. E. Measurement of ²H T₁ and T₁ρ. Relaxation Times in Uniformly ¹³C-Labeled and Fractionally ²H-Labeled Proteins in Solution. *J. Am. Chem. Soc.* **1995**, *117*, 11536–11544.
- (11) Millet, O.; Muhandiram, D. R.; Skrynnikov, N. R.; Kay, L. E. Deuterium Spin Probes of Side-Chain Dynamics in Proteins. 1. Measurement of Five Relaxation Rates per Deuteron in ¹³C-Labeled and Fractionally ²H-Enriched Proteins in Solution. *J. Am. Chem. Soc.* **2002**, *124*, 6439–6448.
- (12) Li, D.-W.; Brüschweiler, R. A Dictionary for Protein Side-Chain Entropies from NMR Order Parameters. *J. Am. Chem. Soc.* **2009**, *131*, 7226–7227.
- (13) Case, D. A. Molecular Dynamics and NMR Spin Relaxation in Proteins. *Acc. Chem. Res.* **2002**, *35*, 325–331.
- (14) Brüschweiler, R. New approaches to the dynamic interpretation and prediction of NMR relaxation data from proteins. *Curr. Opin. Struct. Biol.* **2003**, *13*, 175–183.
- (15) Philippopoulos, M.; Mandel, A. M.; Palmer, A. G.; Lim, C. Structure and dynamics of the M13 coat signal sequence in membranes by multidimensional high-resolution and solid-state NMR spectroscopy. *Proteins: Struct., Funct., Gen.* **1997**, *27*, 481–493.
- (16) Showalter, S. A.; Brüschweiler, R. Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: Application to the AMBER99SB force field. *J. Chem. Theory Comput.* **2007**, *3*, 961–975.
- (17) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct. Bioinform.* **2006**, *65*, 712–725.
- (18) Buck, M. S.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell, A. D. Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme. *Biophys. J. Biophys. Lett.* **2006**, *90*, L36–L38.
- (19) Soares, T.; Daura, X.; Oostenbrink, C.; Smith, L.; van Gunsteren, W. F. Validation of the GROMOS force-field parameter set 45A3 against nuclear magnetic resonance data of hen egg lysozyme. *J. Biomol. NMR* **2004**, *30*, 407–422.
- (20) Lipari, G. A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.* **1982**, *104*, 4546–4559.
- (21) Horita, D. A.; Zhang, W.; Smithgall, T. E.; Gmeiner, W. H.; Byrd, R. A. Dynamics of the Hck-SH3 domain: comparison of experiment with multiple molecular dynamics simulations. *Protein Sci.* **2000**, *9*, 95–103.
- (22) Andrews, B. K.; Romo, T.; Clarage, J. B.; Pettitt, B. M.; Phillips, G. N. Characterizing global substates of myoglobin. *Structure.* **1998**, *6*, 587–594.
- (23) Koller, A. N.; Schwalbe, H.; Gohlke, H. Starting Structure Dependence of NMR Order Parameters Derived from MD Simulations: Implications for Judging Force-Field Quality. *Biophys. J. Biophys. Lett.* **2008**, *95*, L04–L06.
- (24) Smith, P. E.; Pettitt, B. M.; Karplus, M. Stochastic dynamics simulations of the alanine dipeptide using a solvent-modified potential energy surface. *J. Phys. Chem.* **1999**, *97*, 6907–6913.
- (25) Genheden, S.; Ryde, U. A Comparison of Different Initialization Protocols to Obtain Statistically Independent Molecular Dynamics Simulations. *J. Comput. Chem.* In press. DOI: 10.1002/jcc.21546.
- (26) Lawrenz, M.; Baron, P.; McCammon, J. A. Independent-Trajectories Thermodynamic-Integration Free-Energy Changes for Biomolecular Systems: Determinants of H5N1 Avian Influenza Virus Neuraminidase Inhibition by Peramivir. *J. Chem. Theory Comput.* **2009**, *5*, 1106–1116.
- (27) Collins, P. M.; Hidari, K. I. P. J.; Blanchard, H. Slow diffusion of lactose out of galectin-3 crystals monitored by X-ray crystallography: possible implications for ligand-exchange protocols. *Acta Crystallogr., Sect. D* **2007**, *63*, 415–419.
- (28) Houlzelstein, D.; Goncalves, I. R.; Fadden, A. J.; Sidhu, S. S.; Cooper, D. N.; Drickamer, K.; Leffler, H.; Poirer, F. Phylogenetic Analysis of the Vertebrate Galectin Family. *Mol. Biol. Evol.* **2004**, *21*, 1177–1187.
- (29) Leffler, H.; Carlsson, S.; Hedlund, M.; Quian, Y. Introduction to galectins. *Glycoconjugate J.* **2002**, *19*, 433–440.
- (30) Liu, F.-T.; Rabinovich, G. A. Galectins as modulators of tumour progression. *Nat. Rev. Cancer* **2005**, *5*, 29–41.
- (31) Nakahara, S.; Oka, N.; Raz, A. On the role of galectin-3 in cancer apoptosis. *Apoptosis* **2005**, *10*, 267–275.
- (32) Liu, F.-T. Regulatory Roles of Galectins in the Immune Response. *Int. Arch. Allergy Immunol.* **2005**, *136*, 385–400.

- (33) Ilarrgui, J. M.; Bianco, G. A.; Toscano, M. A.; Rabinovich, G. A. New targets III: The coming of age of galectins as immunomodulatory agents: impact of these carbohydrate binding proteins in T cell physiology and chronic inflammatory disorders. *Ann Rheum. Dis.* **2005**, *64*, 96–103.
- (34) Patterson, R. J.; Wang, W.; Wang, J. L. Understanding the biochemical activities of galectin-1 and galectin-3 in the nucleus. *Glycoconj. J.* **2002**, *19*, 499–506.
- (35) Dumic, J.; Dabelic, S.; Flögel, M. Galectin-3: An open-ended story. *Biochim. Biophys. Acta* **2006**, *1760*, 616–635.
- (36) Bachhawat-Sikder, K.; Thomas, C. J.; Suriola, A. Thermodynamic analysis of the binding of galactose and poly-N-acetyllactoseamine derivatives to human galectin-3. *FEBS Lett.* **2001**, *500*, 75–79.
- (37) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, 2008.
- (38) Diehl, C.; Genheden, S.; Modig, K.; Ryde, U.; Akke, M. Conformational entropy changes upon lactose binding to the carbohydrate recognition domain of galectin-3. *J. Biomol. NMR* **2009**, *45*, 157–169.
- (39) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G.; Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–9678.
- (40) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (41) Wu, X.; Brooks, B. R. Self-guided Langevin dynamics simulation method. *Chem. Phys. Lett.* **2003**, *381*, 512–518.
- (42) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (43) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (44) Buck, M.; Karplus, M. Internal and overall peptide group motion in proteins: molecular dynamics simulations for lysozyme compared with results from X-ray and NMR spectroscopy. *J. Am. Chem. Soc.* **1999**, *121*, 9645–9658.
- (45) Zwansig, R.; Ailawadi, N. K. Statistical Error Due to Finite Time Averaging in Computer Experiments. *Phys. Rev.* **1969**, *1982*, 280–282.
- (46) Lu, C.-Y.; Bout, D. A. V. Effect of finite trajectory length on the correlation function analysis of single molecule data. *J. Chem. Phys.* **2006**, *125*, 124701–124709.
- (47) Madsen, H. *Time Series Analysis*; Chapman & Hall/CRC: New York, 2008.
- (48) Efron, B.; Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statist. Sci.* **1986**, *1*, 54–77.
- (49) Prompers, J. J.; Brüschweiler, R. General Framework for Studying the Dynamics of Folded and Nonfolded Proteins by NMR Relaxation Spectroscopy and MD Simulation. *J. Am. Chem. Soc.* **2002**, *124*, 4522–4534.
- (50) Nederveen, A. J.; Bonvin, A. NMR relaxation and internal dynamics of ubiquitin from a 0.2 microsec MD simulation. *J. Chem. Theory Comput.* **2005**, *1*, 363–374.
- (51) Bowers, K.; Devolder, B.; Yin, L.; Kwan, T. A maximum likelihood method for linking particle-in-cell and Monte-Carlo transport simulations. *Comput. Phys. Commun.* **2004**, *164*, 311–317.
- (52) Dempster, A. P.; Laird, N. M.; Rubin, M. D. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. B* **1977**, *39*, 1–38.
- (53) Meiler, J.; Promper, J. J.; Peti, W.; Griesinger, C.; Brüschweiler, R. Model-Free Approach to the Dynamic Interpretation of Residual Dipolar Couplings in Globular Proteins. *J. Am. Chem. Soc.* **2001**, *123*, 6098–6107.
- (54) Umemoto, K.; Leffler, H.; Venot, A.; Valafar, H.; Prestegard, J. H. Conformational differences in liganded and unliganded states of Galectin-3. *Biochemistry* **2003**, *42*, 3688–3695.
- (55) Pfeiffer, S.; Fushman, D.; Cowburn, D. Simulated and NMR-Derived Backbone Dynamics of a Protein with Significant Flexibility: A Comparison of Spectral Densities for the β ARK1 PH Domain. *J. Am. Chem. Soc.* **2001**, *112*, 3021–3036.
- (56) Kanibolotsky, D. S.; Ivanova, O. S.; Lisnyak, V. V. Comparison of NMR and MD NZH bond order parameters: example of HIV-1 protease. *Mol. Sim.* **2006**, *32*, 1155–1163.
- (57) MacRaild, C. A.; Daranas, A. H.; Bronowska, A.; Homans, S. W. Global Changes in Local Protein Dynamics Reduce the Entropic Cost of Carbohydrate Binding in the Arabinose-binding Protein. *J. Mol. Biol.* **2007**, *368*, 822–832.
- (58) Markwick, P. R. L.; Bouvignies, G.; Blackledge, M. S. Exploring Multiple Timescale Motions in Protein GB3 Using Accelerated Molecular Dynamics and NMR Spectroscopy. *J. Am. Chem. Soc.* **2007**, *129*, 4724–4730.
- (59) Maragkis, P.; Lindorff-Larsen, K.; Eastwood, M. P.; Dror, R. O.; Klepeis, J. L.; Arkin, I. T.; Jensen, M. O.; Xu, H.; Trbovis, N.; Friesner, R. A.; Plamer III, A. G.; Shaw, D. E. Microsecond Molecular Dynamics Simulation Shows Effect of Slow Loop Dynamics on Backbone Amide Order Parameters of Proteins. *J. Phys. Chem. B* **2008**, *112*, 6155–6158.
- (60) Tong, Y.; Ji, C. G.; Mei, Y.; Zhang, J. Z. H. Simulation of NMR Data Reveals That Proteins' Local Structures Are Stabilized by Electronic Polarization. *J. Am. Chem. Soc.* **2009**, *131*, 8636–8641.
- (61) Zhang, F.; Brüschweiler, R. Contact Model for the Prediction of NMR N-H Order Parameters in Globular Proteins. *J. Am. Chem. Soc.* **2002**, *124*, 12654–12655.

CT900696Z

JCTC

Journal of Chemical Theory and Computation

Energy Matrix of Structurally Important Side-Chain/ Side-Chain Interactions in Proteins

Karel Berka,[#] Roman A. Laskowski,[‡] Pavel Hobza,^{†,#} and Jiří Vondrášek^{*,†,§}

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, Flemingovo nám. 2, Prague, Czech Republic, Institute of Biotechnology, Academy of Sciences of the Czech Republic, Videnska 1083, 142 00 Prague, Czech Republic, Palacký University, Department of Physical Chemistry, Faculty of Science, t. 17. listopadu 12, 771 46, Olomouc, Czech Republic, and EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Received January 5, 2010

Abstract: The interactions between amino acid side chains in proteins are generally considered to be the most important stabilizing factor controlling the precise arrangement of the polypeptide chain into a well-defined spatial structure. We used the RI-DFT-D method to calculate the full 20×20 matrix of interaction energies between all pairs of amino acid side chains. For each pair, we used a representative 3D conformation extracted from an analysis of known protein structures from Protein Data Bank (PDB). The representative comes from the largest cluster of relative orientations of the two side chains. We find that all of the calculated interaction energies between selected pairs of amino acids are attractive in the gas phase with the exception of side chain pairs having the same total charge. We compared these data with those calculated by the parm03 and OPLS-AA/L force fields to investigate the reliability of simple methods in modeling biomolecules and their behavior. The force fields yield good overall interaction energies for our set but have problems in evaluation of some particular interactions which could be of principal importance for protein stability. We then looked in detail at the 20 side chain interactions involving tryptophan. The histograms of interaction energies showed that the distributions of the interaction energies are neither normal nor Boltzmann-like and that our representative geometries correspond mostly to the minimum energy geometry which is rather poorly populated in the whole pairwise energy distribution. We concluded that cluster representatives obtained by the clusterization algorithm based on geometry criteria cannot be considered as a typical interaction for the whole side chain/side chain interaction distribution. They seem to epitomize the strongest interactions in a protein and are often functionally or structurally important.

Introduction

Proteins are built from 20 natural L-amino acids polymerized into a linear chain of various lengths which, with the exception of the “intrinsically unstructured proteins”, fold into a specific and rigid 3D structure either spontaneously or with the help of various factors (chaperones etc.).¹ Anfinsen’s postulate that protein structure is unambiguously

defined by the amino acid sequence is still, to a large extent, valid.² The polypeptide chain bears specific and heterogeneous chemical properties given by the different nature of composing amino acids. There is a long history of the efforts to collect and analyze the interactions between amino acid side chains in protein structures — mostly determined experimentally by X-ray crystallography or NMR methods. In the past, Miyazawa and Jernigan^{3–5} and others^{6–8} attempted to rationalize the character of the contacts between the side chains and to associate it with contact free energy. Such pairwise contact free energies have proven to be useful for scoring the native folds.⁹ As the number of solved protein structures has become greater, the distance-dependent and orientation statistical potentials have also been proposed.^{9–11} Side-chain/side-chain contacts are characterized geometri-

* Corresponding author tel.: +420 220-410-324, fax: (+420) 220-410-320, e-mail: jiri.vondrasek@uochb.cas.cz.

[†] Academy of Sciences of the Czech Republic.

[‡] European Bioinformatics Institute.

[§] Institute of Biotechnology, Academy of Sciences of the Czech Republic.

[#] Palacký University.

cally and in detail in an online accessible database of side-chain/side-chain interactions created by Laskowski et al.¹²

It is necessary to mention that the predicted free energies calculated from the contact analysis data depend on several approximations, which might not be fully valid for all proteins, as was nicely reviewed by Thomas and Dill.¹³ They examined *a priori* potentials based on a simple hydrophobic-polar model. The calculated energies for all of the possible structures in a two-dimensional lattice resulted in the minimal “native” structure, which helped to construct a new potential recursively. They found that the frequencies of the selected pair of amino acids are not independent in terms of the frequencies of the other amino acids in the context of a sequence and that the extracted potential depends quite remarkably on the chain length and the composition.

To be able to evaluate the free energy of a particular amino acid in a pair interaction, one needs computational methods covering both the enthalpy and entropy terms given by the expression for the Gibbs free energy of association. The achievement of this goal can be significantly complicated by two principal difficulties. First is the level of accuracy for the enthalpy term calculation. The empirical potentials usually utilized are not of the required precision especially when the effect of the solvent has to be taken into account. Second, there is no rigorous and reliable theoretical method to evaluate the entropy term at the same level of accuracy as that for the enthalpy term. Most of the methods for the calculation of the entropic contribution are based on the positional variability determined by the NMR technique.¹⁴

There have been a few attempts to make a comparison of the statistical potential and the *ab initio* calculation of the interaction energy of amino acid side chains. Morozov et al.¹⁵ reported remarkable correspondence between the knowledge-based potential of the hydrogen-bond geometries representing amino acid interactions in proteins and the *ab initio* DFT and MP2 calculations of the hydrogen-bonding energies for model systems. The same authors attempted to evaluate the potential energy surface (PES) for the interaction of aromatic residues at the MP2 and empirical potential levels. The main conclusion of this work is that the interaction is fairly well captured by the empirical potential and “that interactions between cyclic side chains contribute to the geometric distributions observed in protein structures”.¹⁶

Here, we present the results of our study in which we describe and evaluate the interaction energies for all 20×20 amino acid side-chain pairs using representative geometries obtained from analysis of known 3D structures of proteins. We use several force fields as well as quantum chemistry methods both in the gas phase and in a protein/water environment. The importance of the obtained energy values for each interacting side chain pair is discussed in the context of the total interaction energy distribution between amino acid side chains.

Methods

Representative Set Selection. To obtain a representative set of amino acid side-chain pairs, we extracted data from a

specially updated version of the Atlas of Protein Side-Chain Interactions from October 2006 (<http://www.ebi.ac.uk/thornton-srv/databases/sidechains>). The Web atlas is based on the printed atlas published in 1992 by Singh and Thornton.¹⁷ It analyzes the interaction geometries of all 20×20 amino acid side-chain pairs as derived from a nonhomologous data set of 2548 3D structural models of proteins solved by X-ray crystallography to a resolution of 2.0 Å or better. For each of the 20×20 pairs of side chain types, each distance of side chain 2 interacting with side chain 1 is transformed into a common reference frame defined by side chain 1.

The preferred interaction geometries are determined from the local clustering in 3D of the distribution of side chains 2 relative to side chain 1. For each cluster, the most representative side chain 2 is selected, being the side chain which has the minimum total root mean squared distance to all of the other side chains in the cluster. A more detailed description can be found in ref 18. In the work described here, we used the cluster representative from the largest cluster in each of the 20×20 distributions. Figure 1 shows top clusters, and their representative side chains, for four example distributions, each involving Trp as side chain 1. Figure 1a and b show the top cluster geometries for Asp and Ser, respectively. Here, the location of side chain 2 is such that it can form a hydrogen bond with the nitrogen of the tryptophan. In Figure 1c and d, the interacting side chains are Leu and Lys. Here, the interactions are hydrophobic in nature, and consequently less specific and less directional.

Geometry Preparation. Each residue was truncated at the C α atom of the protein backbone, and hydrogen atoms were added using a modified side-chain only force field¹⁸ implemented in the Gromacs molecular dynamics package.¹⁹ It means for example that glycine is approximated by CH₄ and Alanine by C₂H₆ groups. All of the possible positions of the polar hydrogens of the merkaptyl and hydroxyl groups of Cys, Ser, Thr, and Tyr were generated along with two neutral isomers of histidine. Proline was modeled as a cyclic tetrahydropyrrole. This model captures all specific features of proline interactions (pseudoplanarity, cyclic structure) as was shown by Biedermannova et al.²⁰ The positions of the hydrogens were then optimized in complex geometry for each pair using the SCC-DFTB-D method²¹ in the DFTB+ package.²² The hydrogens in the pairs containing at least one charged residue were optimized separately. The most stable pair determined by means of the benchmark method (see below) was then used for further calculations.

Calculation of the Gas-Phase Interaction Energies. The quantum mechanical energies were calculated using the RI-DFT-D/TPSS/TZVP method.²³ The RI-DFT-D energies were calculated with the Turbomole 5.9 package.²⁴ This BSSE-free method has proved to be reasonably accurate and computationally efficient on the subset of geometries calculated previously.¹⁸

We also used two modified force fields parametrized earlier – OPLS-AA/L²⁵ and parm03.²⁶ These force fields contain only amino acids truncated at the C α atom. The residual nonintegral charge is further distributed over added hydrogen atoms attached to the C α atom. The noncovalent

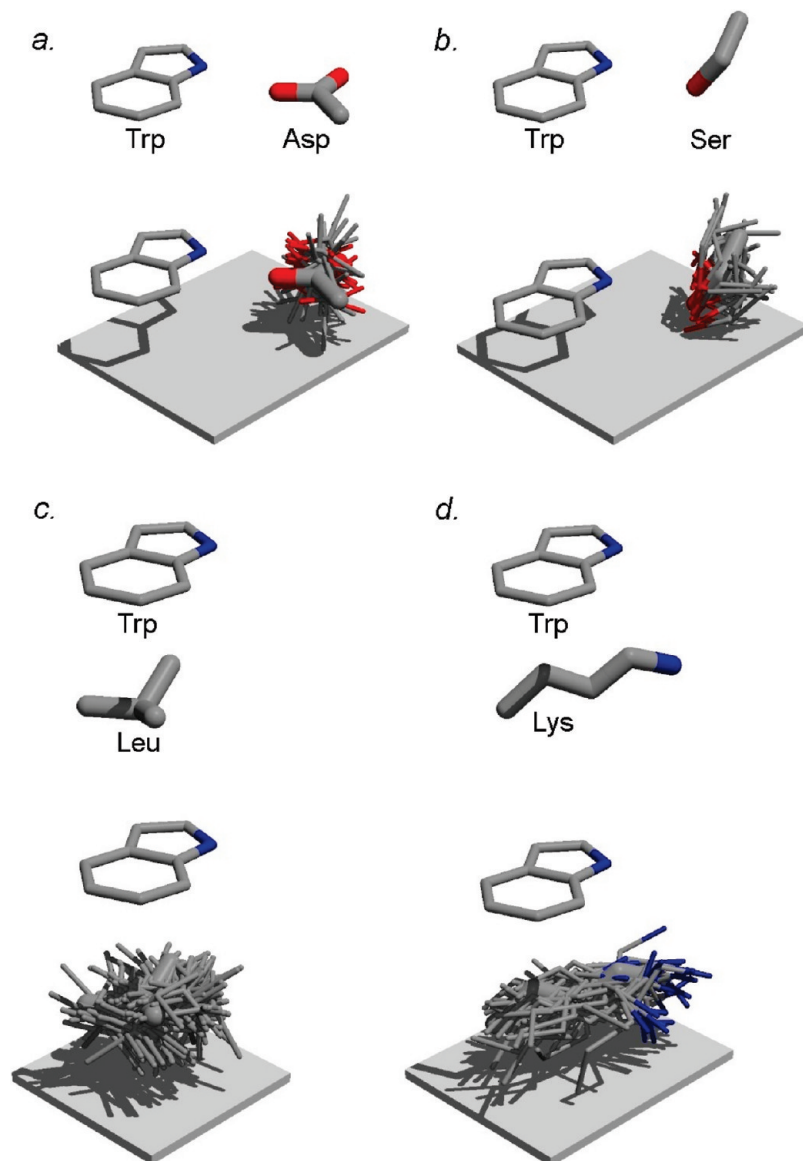


Figure 1. Some examples of side chain interactions in protein 3D structures. All examples involve interactions with tryptophan. The side chains shown are (a) aspartic acid, (b) serine, (c) leucine, and (d) lysine. Each diagram consists of two parts. The lower part shows the largest cluster of the interacting side chains, as extracted from a representative data set of protein structures in the PDB. The “cluster representative” is shown with thicker bonds. This corresponds to the side chain with the lowest total distance to all the other members of the cluster. In the upper part of each figure is shown just the Trp side chain and the cluster representative, both labeled by their three-letter code. The figure was rendered using Raster3D.³⁵

interactions were calculated as a sum of the electrostatic and Lennard-Jones terms for the complexes of amino acid fragments forming a particular pair. The force-field calculations were performed with the Gromacs 4.0 package.²⁷

Solvent Effect. The effect of an environment was evaluated by the RI-DFT-D method utilizing the COSMO model implemented in the Turbomole package.²⁸ Two dielectric constants were used to model the effect of a protein/water environment ($\epsilon = 4, 80$) on the interaction energies.

Results

Benchmark Energy Calculations—Gas Phase Interaction Energies. We have previously shown¹⁸ a correlation between interaction energies for the selected set of side chains evaluated by means of various theoretical methods

(CCSD(T)|CBS, DFT-D, RI-DFT-D, MP2, OPLS-AA/L, and parm03 force field). We have found that RI-DFT-D is a reasonable compromise between the accuracy and the speed of the calculations, which supports our choice of this method as the reference. In this paper, we expanded the set to all of the possible combinations of 20×20 amino acid side-chain pairs. All of the geometries of the calculated pairs were selected by the cluster analysis described above to represent significantly populated geometry arrangements of interacting amino acids. The reference interaction energies for these pairs calculated by the RI-DFT-D method thus represent a measure of affinity based on the positions of the side chains determined experimentally and stored in the PDB database.²⁹ The final numbers are presented in Table 1.

Table 1. Gas Phase Interaction Energy Matrix for the Cluster Representatives for All of the 20 × 20 Possible Pairs between Residues within Proteins Calculated with the RI-DFT-D/TPSSITZVP Method (All energies are in kcal/mol)

DFTD	G	A	V	I	L	F	Y	W	H	P	T	S	N	Q	C	M	K	R	D	E
G	-0.6	-0.7	-0.8	-0.8	-0.9	-1.0	-0.8	-1.6	-0.9	-0.2	-0.9	-1.0	-0.8	-0.8	-1.0	-0.9	-1.8	-0.4	-1.6	-3.8
A	-0.3	-0.2	-1.0	-1.4	-1.3	-1.7	-2.1	-0.7	-1.2	-1.2	-1.2	-1.4	-1.7	-1.5	-0.6	-1.5	-2.4	-3.3	-3.0	-4.6
V	-0.9	-1.5	-1.8	-1.8	-1.3	-1.3	-1.4	-2.1	-0.8	-2.1	-1.1	-1.8	-1.1	-1.5	-0.9	-1.1	-3.5	-3.4	-3.9	-2.9
I	-1.1	-1.5	-1.2	-1.5	-1.7	-3.0	-1.5	-3.0	-1.1	-1.2	-1.2	-1.7	-1.3	-1.6	-0.6	-0.7	-3.8	-3.4	-4.8	-3.3
L	-1.0	-1.0	-1.3	-1.5	-2.0	-2.4	-1.8	-3.9	-1.9	-2.3	-1.4	-1.6	-1.7	-2.3	-1.1	-2.0	-4.9	-4.5	-6.0	-6.4
F	-0.8	-1.4	-1.9	-2.7	-2.3	-2.1	-2.2	-4.6	-2.6	-2.8	-2.5	-2.5	-4.3	-3.0	-1.1	-2.1	-5.7	-9.0	-10.2	-10.2
Y	-0.7	-1.3	-2.5	-2.9	-2.3	-3.3	-3.7	-5.3	-2.8	-4.0	-1.9	-2.9	-3.4	-3.7	-1.3	-2.6	-8.1	-10.4	-29.5	-34.6
W	-1.8	-1.9	-4.5	-2.5	-2.5	-6.0	-5.6	-4.9	-4.5	-3.4	-7.4	-7.0	-5.2	-5.1	-3.7	-4.9	-9.0	-12.6	-27.4	-27.6
H	-0.9	-1.7	-1.7	-3.0	-2.7	-3.0	-2.8	-5.4	-6.1	-2.4	-5.4	-6.1	-8.3	-7.4	-3.0	-1.9	-6.8	-7.7	-27.8	-24.3
P	-1.2	-1.2	-1.9	-1.6	-1.9	-3.3	-1.6	-4.1	-3.4	-1.7	-2.4	-1.7	-0.8	-1.8	-1.1	-1.9	-1.8	-2.9	-6.5	-5.5
T	-0.5	-1.2	-0.2	-1.2	-1.2	-1.0	-3.1	-7.8	-8.0	-0.7	-7.2	-1.7	-2.5	-2.2	-0.8	-1.5	-1.7	-16.9	-12.7	-12.8
S	-0.4	-0.9	-1.8	-1.3	-1.5	-1.4	-2.3	-2.7	-0.3	-2.2	-7.3	-2.9	-6.7	-2.1	-1.1	-2.0	-9.0	-16.0	-13.8	-10.8
N	-0.9	-1.2	-1.3	-0.8	-2.1	-2.8	-3.9	-4.0	-2.6	-1.7	-0.9	-6.6	-7.2	-5.5	-1.8	-2.2	-29.8	-21.4	-25.8	-25.9
Q	-1.1	-1.3	-1.4	-1.7	-1.5	-2.1	-2.1	-4.5	-3.6	-2.7	-2.6	-1.9	-7.1	-9.8	-1.7	-2.3	-6.2	-20.9	-24.3	-25.5
C	-0.6	-0.5	-0.9	-0.7	-1.3	-1.6	-1.3	-3.6	-4.1	-1.1	-0.8	-0.9	-2.2	-2.4	-59.9	-2.4	-5.7	-9.8	-10.6	-8.8
M	-1.2	-0.5	-1.1	-1.4	-2.2	-2.7	-2.4	-2.5	-0.7	-0.9	-1.3	-1.6	-3.8	-3.0	-1.4	-1.9	-6.4	-7.9	-7.0	-11.9
K	-1.9	-2.2	-3.8	-3.7	-3.1	-5.7	-9.5	-6.8	-3.8	-1.3	-2.1	-7.1	-28.9	-28.3	-5.5	-7.3	58.7	55.8	-113.8	-113.7
R	-1.6	-2.8	-3.6	-3.8	-3.6	-7.5	-8.6	-10.6	-6.1	-1.4	-3.5	-15.7	-20.0	-22.8	-5.7	-7.5	51.1	50.7	-115.6	-107.1
D	-1.4	-3.1	-3.3	-5.7	-6.0	-7.1	-40.1	-24.1	-31.6	-8.7	-7.0	-12.0	-27.1	-26.8	-6.7	-4.2	-116.1	-126.5	62.5	50.1
E	-2.1	-2.8	-3.7	-4.1	-4.5	-4.9	-37.2	-27.2	-26.6	-8.2	-12.0	-12.5	-7.2	-26.0	-9.0	-8.6	-109.9	-140.1	51.9	70.4

The first of the important results is that all of the interaction energies for structure representatives presented in Table 1 are attractive with only a few regular exceptions—the pairs containing amino acids with a similar charge. The result thus reflects an important fact regarding the protein's intramolecular stabilization provided by the selective arrangement of interacting amino acids. It must be stressed again that the interaction energy of all of the amino acid pairs was calculated exclusively for the most populated cluster representative geometry, which most probably represents the local distance minimum. The lack of destabilizing contributions is then not so surprising. We can imagine the existence of sterical barriers caused by a tight arrangement of secondary structure elements which could include the studied amino acids. Such an environment may sometimes push the amino acids out of the attractive regime and could result in repulsive behavior of the interacting amino acids. We do not report a single case of such an interaction mode (with the above-mentioned exceptions) for the studied set.

Asymmetry of the Interaction Energy Matrix. The second of the important results which should be properly explained is the asymmetry of the interaction energy matrix. The asymmetry of the matrix is a consequence of the way the clusters were calculated. Figure 2 shows the explanation of the matrix feature. In Figure 2, a and b show two separate interactions between side chains of type S and T: S1 with T1 and S2 with T2. When these interactions are superposed in the frame of reference of the S residues (c), the T residues come close together and might fall in the same cluster. However, when the interactions are superposed on the T residues (d), the S residues are thrown apart and would be unlikely to fall into the same cluster. For the side chain Atlas, this is not a problem, as one is interested in the distribution of side chain B around side chain A, and where the highest concentrations of B are, relative to A (and vice versa). For a symmetrical matrix, however, one would need to calculate

full distributions for each pair of amino acids. So, in Figure 2c, one would need to calculate the RMSD between T1 and T2 and to add to it the RMSD between S1 and S2 when the T side chains are superposed (as in d). We still think that the definition used in the Atlas is a fair way to look at the data and accept the asymmetry because we are interested, in principle, about significant structure features which are most probably based on certain geometry preferences between interacting amino acids.

It is worth stressing here again that first we have calculated the pairwise noncovalent interactions between amino acids in the gas phase and that the influence of the environment has not been taken into account. The fact that the system exists in an environment can change the stabilization

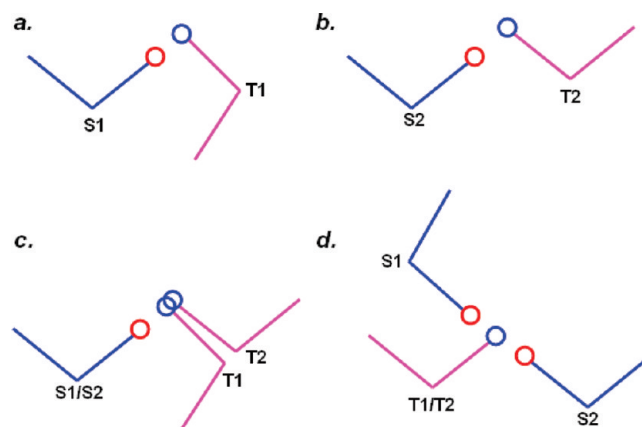


Figure 2. (a and b) Two separate interactions between side chains of type S and T: S1 with T1 and S2 with T2. When these interactions are superposed in the frame of reference of the S residues (c), the T residues come close together and might fall in the same cluster. However, when the interactions are superposed on the T residues (d), the S residues are thrown apart and would be unlikely to fall into the same cluster.

proportions considerably is part of the latter chapter about solvent effects.

We divided the interactions into several groups according to the chemical properties of the compositional amino acids, and in the following paragraphs we shall describe the results separately.

Charge–Charge. The most attractive interactions in the gas phase were obtained for salt bridges — approximately 100 kcal/mol. The interaction energies where Arg is in a pair with a negatively charged carboxylic acid are slightly stronger (up to 140 kcal/mol) in comparison with those containing a Lys residue. This may be the effect of the additional hydrogen bond from the second NH group participating in the interaction or the influence of the electron distribution of the guanidinium group.

Charge–Neutral. The interactions of the charged residues are generally quite strong in the complexes with polar or aromatic residues — around 10–20 kcal/mol for positively charged and 20–30 kcal/mol for negatively charged residues. This difference is more profound for the charged-aromatic pairs. The geometry of the negatively charged-aromatic pairs is different from that of those containing positively charged side chains. The basic amino acids usually interact with aromatics in a stacking-like manner, unlike the acidic residues, which prefer more directional, mostly H-bond interaction. We can find a further difference between Arg and Lys which arises from the fact that Arg mostly stacks above the plane of the ring of the aromatics, indicating clearly a more dispersive character of interaction. Lys, on the other hand, possesses its long aliphatic chain above the ring, so the charged amine group is farther from the ideal contact with the aromatic ring. Both negatively charged residues are oriented in such a way that their carboxylic group is in the plane of the ring with negligible electron contacts with the aromatics. They interact either with the hydroxy group of Tyr or with the amide group of Trp or His. In the case of Glu, Asp–Phe interactions, the carboxylic group is to a certain extent in an interaction with the main chain of the aromatic and is not oriented above the highest electron density on the aromatics.

Polar–Polar and Polar–Aromatics. The third class of interactions is polar–polar and polar–aromatic contacts. Their interaction energy is about 5 kcal/mol. As they are mostly based on the formation of hydrogen bonds in an orientation-dependent manner, the resulting interaction energies fluctuate in the largest range even for the same pairing of amino acids. Good examples are the Thr–Ser pair, where the interaction energy is only –1.7 kcal/mol, and the Ser–Thr pair, which is much stronger, namely at –7.3 kcal/mol. At this point, we have to stress that in both cases the best combination of the rotamers and optimal position of both of the hydroxyl groups was used.

Aromatics–Aromatics. It is well-known that the aromatic–aromatic pairing is abundant in proteins and is also quite homogeneous because of the similar character of the interacting residues. Their interaction energies on average are around 5 kcal/mol. The strongest interaction among aromatic residues comes from pairs containing Trp, mostly due to the aromatic character and size of its indole ring. The

Trp is followed by His, which interacts mostly through the hydrogen bond. It should also be noted that the result is dramatically influenced by the selection of an appropriate isomer.

Aliphatics–Others. The largest group of interactions comprises pairs containing aliphatic residues. Their interaction energy with most of the residues is quite small (below 2 kcal/mol) with the exception of the aliphatic–charged and the aliphatic–aromatic pairs, which are stronger. Polar residues cannot create hydrogen bonds and are perceived mainly by dispersion interactions. Proline exhibits special features: it behaves similarly to an aliphatic residue of the same size (Leu, Ile), but its interaction with the charged residues is different.

Sulfurics–Others. A small group can also be derived from sulfur-atom-containing residues, namely, Cys and Met. Their interaction energy is similar to aliphatic residues but with several notable deviations. The biggest difference is the Cys–Cys pair, which is bound covalently and thus cannot be compared to the other cases. However, its dissociation energy for the disulfide bond is about 65 kcal/mol.³⁰ Because of the better polarizability of the sulfur atom, its interaction with charged residues is approximately twice as strong as in the case of the charged-aliphatic pairs.

Looking at each residue individually in terms of its total interaction energy with all amino acids, we can create a “stability line”. It runs from the residue of the highest stabilization potential to the lowest. The energy differences between adjacent residues in the stability line are not constant, so they can be grouped into subclasses. The “>” signifies an important change in the interaction energies:

$$D, E > R > K > N, Q > W, Y > H > S > T > F > \\ M, C > P, L > I, V > A > G$$

The strongest stabilization not surprisingly comes from interactions of charged residues even when we take into account the repulsion between amino acids of the same charges. The stability line continues with polar and aromatic residues and ends with aliphatic residues according to their size. It should be mentioned again here that these values are gas phase interaction energies, and hence no effect of an environment has been taken into account.

Parm03 and OPLS-AA/L Force Field Interaction Energies. We evaluated the interaction matrix for the same set of structures with two force fields typically used for the protein study, i.e., Amber parm03 and OPLS-AA/L. We aimed to provide a quantitative comparison between the results obtained by the RI-DFT-D energies and the molecular-mechanical force-field methods. One has to be aware of the difference between the calculation of interaction energies by empirical study and by the quantum chemical approach. In the empirical potential case, the interaction energy is calculated as a sum of the nonbonded interactions between each atom in both residues. The interactions involve a Coulombic term for electrostatics and the Lennard-Jones 6–12 term covering the van der Waals contributions. In the case of RI-DFT-D, the interaction energy is calculated as the difference between the total energy of the complex and the energies of both subsystems.

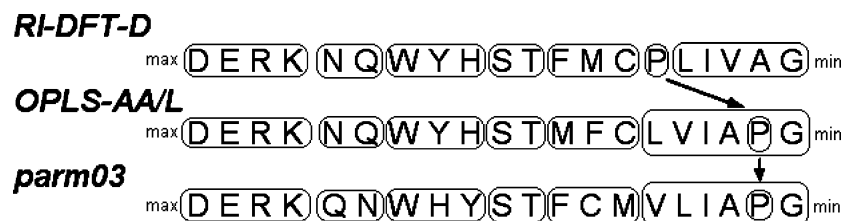


Figure 3. Amino acid families sorted by their summed interaction energy for RI-DFT-D, parm03, and OPLS calculations.

The most notable differences between the force-field and RI-DFT-D energies are substantially weaker interactions per residue provided by empirical potential methods. In contrast with the RI-DFT-D results, we detected additional repulsive behavior of some interacting pairs with the exception of the pairs with the same charges. One of the reasons may be accounted for the Lennard-Jones repulsive term in the force field being generally too steep. This can cause difficulties arising from the fact that the positions of the hydrogens were optimized at the RI-DFT-D level, which consequently shortens the inter-residue distance. This usually leads to an increase of the Coulombic term (repulsive for two hydrogen atoms) and also to the enlargement of the repulsion coming from the Lennard-Jones term in the force-field calculation. Both these aspects can contribute to the overall repulsion calculated energy computed by the force-field method for the pairs which are still attractive in RI-DFT-D. We can also rationalize a reason for a higher number of repulsions coming from the Pro residue from the way we treated this amino acid. The simplification of Pro alters the hybridization state of the N atom from sp^2 (planar) to sp^3 (tetrahedral). This situation improved the interaction energy in quantum chemical methods in contrast with force fields, which lacks proper parameters for such a state.

Last but not least, the partial charges on the amino acids in the force field are generally adjusted for the solvent environment. Depending on the geometry of the interacting amino acid in the structural context of the protein, these interactions are sometimes under/overestimated, which also strengthens the difference in comparison with the quantum chemical interaction calculations in the gas phase.

The previously defined “stability line” can be further subdivided into “families” of amino acids according to their physical-chemical similarities (Figure 3). We define the families as follows: charged residues (DERK), polar residues with peptide-bond motifs (NQ), aromatic residues with at least one polar atom (WYH), hydroxyl-containing polar residues (ST), polarizable residues with electron-rich regions (FMC), unique proline ring (P), and aliphatic residues sorted by their respective size (LIVAG). As is apparent from Figure 3, the amino acids in families behave similarly in all of the methods used.

Both interaction matrices are similar (data not shown), with their correlation coefficients being higher than 0.95. The differences in the results can be seen at the level of the average interaction energies in comparison with the RI-DFT-D values. Both force fields have lower average values of interaction energies (parm03, -3.8 kcal/mol; OPLS, -4.5 kcal/mol; while RI-DFT-D, -6.2 kcal/mol) as well as median values (parm03, -1.6 kcal/mol; OPLS, -1.6 kcal/mol; while

RI-DFT-D, -2.5 kcal/mol). Particularly, the median values demonstrate that the force fields have a high level of similarity and that the energies are weaker than those obtained by the RI-DFT-D method. An important detail contributing to the difference between the RI-DFT-D and empirical potential results arises from the fact that force-field parameters in both utilized empirical methods are optimized for molecules in a solvent environment.

The Effect of Environment on the Interaction Energies in the Matrix. While all of the interactions in the gas phase can be calculated explicitly and in principle with reasonable accuracy, most of the interactions of biomolecules and their complexes are realized in a protein or water environment, which makes a precise evaluation of the interaction energy complicated if not impossible because of the heterogeneous conditions around the interacting residues. In order to take the environment roughly into account, we used solvent-implicit models. We used two dielectric constants: $\epsilon = 4$, mimicking the effect of a protein environment, and $\epsilon = 80$, for the effect of water. We calculated the interaction energies by the RI-DFT-D method with the COSMO implicit-solvent model.

The results presented in Tables 2 and 3 show that the higher the dielectric constant of the surrounding, the smaller the differences between the interaction energies for all of the interacting pairs of amino acids. The apparent reason is the dielectric screening of the dominant electrostatic interaction. The consequence of this effect is a decrease of the average and the median of the interaction energy. In comparison with the gas phase interaction energy median (-2.5 kcal/mol), the value in a protein-like environment ($\epsilon = 4$) is -1.4 kcal/mol and in a water-like environment ($\epsilon = 80$) is only -0.9 kcal/mol.

Charge–Charge. Charged pairs lose most of their interaction energies upon the introduction of the solvent in comparison with other pairs. This is caused by the screening of a substantial part of their interaction energy being dominated by electrostatics. On the basis of the values presented in Tables 1–3, we observed that the like-charged pairs dropped more in their repulsive interaction energy (33%, 4.4% of interaction in gas phase for $\epsilon = 4, 80$) than the salt-bridge pairs (37%, 7.7% for $\epsilon = 4, 80$ of the gas-phase values). The repulsion existing in the gas phase can even be surpassed, and the pairs of like-charge residues show an attractive character in a water environment (Arg–Arg). This behavior has recently been reported by Vondrášek et al.³¹ on Arg–Arg as a potential stabilizing factor in proteins.

Charge–Neutral. Also, charge–neutral pairs are quite weakened by the presence of a solvent. However, the weakening of the interaction energies is smaller than in the

Table 2. Interaction-Energy Matrix for the Cluster Representatives for All of the 20 × 20 Possible Pairs between Residues Calculated with the RI-DFT-D/TPSSITZVP Method with the COSMO Model in a Protein-Like Environment ($\epsilon = 4$) (All energies are in kcal/mol)

	G	A	V	I	L	F	Y	W	H	P	T	S	N	Q	C	M	K	R	D	E
G	-0.5	-0.7	-0.8	-0.8	-0.8	-0.1	-0.5	-1.1	-0.7	-0.2	-0.7	-0.3	0.1	-0.7	-0.8	-0.7	-1.0	0.3	0.1	-0.4
A	-0.3	-0.2	-1.0	-1.4	-1.3	-1.5	-1.7	-0.2	-1.0	-1.1	0.1	-0.6	-0.7	-1.2	-0.2	-1.3	-0.6	-0.8	-0.6	-0.6
V	-0.8	-1.4	-1.7	-1.7	-1.3	-1.0	-1.1	-1.9	-0.4	-2.0	-1.0	-0.5	-0.7	-1.1	-0.4	-0.8	-1.0	-2.0	-1.2	-0.7
I	-1.0	-1.5	-1.1	-1.4	-1.7	-2.7	-1.4	-2.6	-0.8	-1.2	-1.2	-0.8	-0.8	-1.1	-0.1	-0.4	-1.2	-2.1	-0.9	-0.8
L	-1.0	-1.0	-1.3	-1.5	-1.9	-2.0	-1.6	-3.3	-1.5	-1.4	-1.3	-1.4	-1.1	-1.0	-0.6	-1.5	-2.4	-2.7	-1.6	-1.2
F	-0.6	-1.1	-1.6	-2.3	-1.8	-1.6	-1.9	-3.8	-1.8	-2.4	-1.5	-1.5	-2.5	-2.0	-0.4	-1.3	-2.4	-4.3	-2.4	-1.8
Y	-0.4	-0.9	-2.0	-2.3	-1.7	-2.9	-3.2	-4.3	-1.0	-2.7	-1.2	-1.9	-1.8	-2.3	-0.3	-1.9	-4.4	-4.5	-14.9	-20.4
W	-1.3	-1.4	-3.8	-2.3	-2.0	-4.7	-4.4	-4.0	-3.6	-2.8	-5.3	-4.8	-2.9	-3.4	-2.6	-4.1	-4.0	-5.8	-11.8	-13.9
H	-0.6	-1.3	-1.4	-2.5	-2.0	-2.1	-2.0	-3.6	-2.5	-2.0	-3.7	-4.1	-5.1	-4.8	-1.7	-0.8	-3.2	-3.2	-12.9	-11.0
P	-0.9	-1.1	-1.8	-0.8	-1.2	-2.0	0.1	-2.5	-2.0	-1.4	-0.9	-0.4	0.7	-0.8	-0.6	-0.2	-1.5	-1.9	-1.3	-0.2
T	0.3	-0.3	-0.2	-1.1	-1.1	-0.9	-2.3	-5.3	-5.7	-0.7	-5.2	-0.1	-0.1	-0.6	-0.3	-1.2	-0.6	-9.1	-4.9	-5.2
S	0.5	0.1	-1.3	-1.2	-0.8	-1.1	-1.6	-1.9	-0.1	-1.2	-5.0	-1.2	-4.6	0.0	-0.4	-1.2	-3.0	-8.3	-5.6	-3.7
N	-0.2	-0.6	-0.7	-0.3	-1.3	-1.2	-2.2	-1.8	-1.3	-1.0	-0.6	-4.3	-4.7	-3.0	-0.2	-1.1	-14.7	-10.6	-11.7	-11.9
Q	-0.8	-0.9	-1.0	-1.2	-1.0	-1.2	-1.3	-2.9	-1.9	-2.0	0.1	0.6	-4.6	-7.0	-0.1	-1.4	-2.6	-10.4	-10.8	-12.6
C	-0.4	-0.3	-0.3	-0.2	-0.7	-0.8	-0.5	-2.4	-2.6	-0.7	-0.3	1.4	-0.3	-1.3	-56.3	-1.6	-2.3	-3.3	-2.8	-2.7
M	-0.8	-0.4	-0.9	-1.3	-1.8	-2.1	-1.8	-2.3	0.0	-0.5	-0.9	-1.2	-1.9	-1.8	-0.8	-1.6	-2.5	-2.8	-1.0	-4.0
K	-1.0	-1.4	-1.4	-0.8	-2.0	-2.9	-3.9	-3.0	-1.8	0.4	-0.2	-2.0	-13.5	-13.8	-1.2	-3.1	20.8	19.0	-41.1	-42.1
R	-0.6	-1.7	-2.1	-2.3	-1.9	-2.8	-4.8	-4.9	-3.1	0.3	-2.0	-8.2	-10.2	-11.4	-2.7	-3.4	15.2	15.9	-44.5	-38.3
D	0.4	-0.3	-1.1	-1.9	-1.2	-2.6	-24.9	-11.2	-14.5	-2.7	-1.1	-5.1	-12.1	-12.3	-2.2	-1.4	-41.6	-51.8	23.9	16.3
E	0.5	-0.5	-1.2	-1.1	-0.9	-0.8	-22.9	-13.8	-13.0	-3.0	-4.2	-5.1	-2.5	-13.1	-3.3	-2.8	-38.5	-61.0	16.6	30.6

Table 3. Interaction-Energy Matrix for the Cluster Representatives for All of the 20 × 20 Possible Pairs between Residues Calculated with the RI-DFT-D/TPSSITZVP Method with the COSMO Model in a Water Environment ($\epsilon = 80$) (All energies are in kcal/mol)

	G	A	V	I	L	F	Y	W	H	P	T	S	N	Q	C	M	K	R	D	E
G	-0.5	-0.6	-0.8	-0.7	-0.8	0.0	-0.3	-0.8	-0.5	-0.1	-0.6	0.1	0.7	-0.6	-0.6	-0.6	-0.7	0.5	0.8	1.0
A	-0.2	-0.2	-0.9	-1.3	-1.2	-1.3	-1.6	0.1	-0.9	-1.1	0.9	-0.2	0.0	-1.0	0.0	-1.2	0.0	0.0	0.3	1.0
V	-0.8	-1.4	-1.7	-1.7	-1.3	-0.9	-1.0	-1.7	-0.1	-2.0	-1.0	0.2	-0.4	-0.9	-0.1	-0.6	-0.2	-1.6	-0.4	0.1
I	-1.0	-1.4	-1.1	-1.4	-1.6	-2.5	-1.3	-2.4	-0.6	-1.1	-1.1	-0.2	-0.5	-0.8	0.3	-0.3	-0.4	-1.7	0.4	0.0
L	-0.9	-1.0	-1.2	-1.4	-1.9	-1.8	-1.5	-2.8	-1.2	-0.8	-1.3	-1.3	-0.8	-0.2	-0.4	-1.3	-1.7	-2.2	0.0	0.9
F	-0.4	-0.9	-1.4	-2.0	-1.6	-1.2	-1.7	-3.3	-1.4	-2.1	-0.9	-0.9	-1.3	-1.4	0.0	-0.8	-1.4	-2.6	0.5	1.8
Y	-0.2	-0.6	-1.7	-1.9	-1.4	-2.6	-3.0	-3.7	0.2	-2.0	-0.7	-1.3	-0.7	-1.4	0.3	-1.5	-3.0	-2.0	-8.8	-14.4
W	-0.9	-1.1	-3.4	-2.1	-1.6	-4.0	-3.7	-3.5	-3.1	-2.3	-4.1	-3.6	-1.3	-2.3	-1.9	-3.6	-2.5	-3.1	-4.8	-7.7
H	-0.3	-1.0	-1.2	-2.2	-1.5	-1.5	-1.5	-2.4	0.0	-1.8	-2.6	-2.8	-3.0	-3.0	-0.8	-0.1	-2.1	-1.4	-5.5	-4.5
P	-0.8	-1.1	-1.8	-0.2	-0.7	-1.1	1.2	-1.5	-1.1	-1.3	0.0	0.4	1.6	-0.2	-0.2	0.9	-1.7	-1.9	0.7	1.9
T	0.7	0.3	-0.2	-1.1	-1.1	-0.8	-1.8	-3.9	-4.4	-0.7	-4.2	0.8	1.3	0.2	0.0	-1.1	-0.5	-5.4	-1.5	-1.7
S	1.0	0.6	-1.0	-1.2	-0.4	-0.9	-1.2	-1.3	0.0	-0.7	-3.8	-0.2	-3.5	1.3	0.0	-0.8	-0.2	-4.7	-1.9	-0.3
N	0.3	-0.2	-0.4	0.0	-0.7	-0.1	-1.1	-0.4	-0.4	-0.7	-0.5	-2.9	-2.9	-1.4	0.9	-0.3	-7.4	-5.1	-4.6	-4.8
Q	-0.5	-0.7	-0.8	-0.9	-0.6	-0.6	-0.9	-2.1	-0.7	-1.6	1.8	2.0	-2.9	-5.2	0.9	-0.7	-1.2	-5.2	-4.0	-6.3
C	-0.3	-0.2	0.0	0.2	-0.4	-0.3	0.1	-1.8	-1.7	-0.4	0.0	2.7	1.0	-0.6	-55.6	-1.1	-0.7	-0.2	1.0	0.1
M	-0.6	-0.2	-0.8	-1.2	-1.6	-1.8	-1.4	-2.2	0.5	-0.2	-0.7	-0.9	-0.8	-0.9	-0.4	-1.3	-0.9	-0.6	1.5	-0.5
K	-0.7	-1.2	-0.6	0.1	-1.7	-2.1	-1.5	-2.2	-1.3	0.8	0.2	0.4	-5.8	-6.7	0.8	-1.2	3.0	2.1	-8.1	-9.6
R	-0.3	-1.3	-1.7	-1.9	-1.4	-1.1	-3.5	-2.8	-1.8	0.9	-1.4	-4.6	-5.3	-5.7	-1.4	-1.6	-1.2	0.1	-12.9	-7.4
D	1.1	0.9	-0.5	-0.5	0.5	-1.1	-18.5	-5.5	-5.9	-0.2	1.5	-2.1	-4.4	-5.1	-0.7	-0.3	-7.7	-18.4	6.3	1.2
E	1.5	0.3	-0.4	-0.1	0.4	0.7	-16.7	-7.8	-6.4	-0.7	-0.7	-1.8	0.1	-6.6	-0.5	-0.4	-6.1	-25.2	0.9	12.8

case of charge–charge pairs (43%, 28% for $\epsilon = 4$, 80 of the gas-phase values). This fact is in concord with the smaller total interaction energies, not as dominated by electrostatics as is the case of charge–charge pairs.

Polar–Polar or Aromatics. A solvent has a smaller effect on these pairs in comparison with the previous cases. While pairs with a mixed character of polar and aromatic residues are more sensitive to the effect of a water environment (64%, 41% for $\epsilon = 4$, 80), the polar–polar contacts surprisingly are less affected (65%, 54% for $\epsilon = 4$, 80).

Aromatics–Aromatics. The decreasing sensitivity of these interacting pairs to the effect of the environment is demonstrated by a moderate decrease of the interaction strength (79%, 68% for $\epsilon = 4$, 80). The reason for such insensitivity is the different nature of their interaction as reported in Berka et al.³²

Aliphatics–Others. Aliphatic residues are the least sensitive to the effect of the environment. Their interaction energies are almost constant for aliphatic–aliphatic pairs (95%, 91% for $\epsilon = 4$, 80). Their interactions with polar or

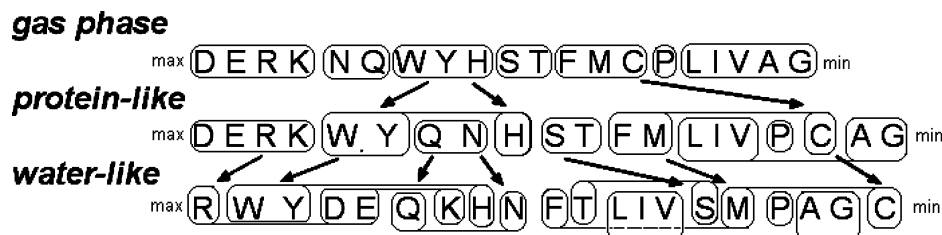


Figure 4. Amino acid families in the environment sorted by their total interaction energy provided by the RI-DFT-D calculations. The dispersion-bound residues are generally shifted upward unlike the electrostatic ones.

aromatic residues are slightly more sensitive (polar: 70%, 59%; aromatic: 79%, 67%), reflecting the different proportionality of the stabilizing forces.

Sulfurics—Others. Sulfur-containing residues act similarly in an environment to the polar residues (62%, 58% for $\epsilon = 4, 80$).

The pair interaction energies between the residues are influenced differently by the solvent depending on their characteristics. Charged residues are the most sensitive to the effect of their environment, followed by polar and sulfur-containing residues. Aromatic and aliphatic residues sustain most of their interaction energy despite the significant environment change. This different sensitivity to the environment changes the positions of amino acids and families in the stability line quite significantly, as can be seen in Figure 4.

The strongest effect of the environment is a change of the relative positions of residues in the stability line. The environment promotes the interaction between the residues of an aromatic or aliphatic character (mainly Trp, Tyr, Leu, Ile, and Val). On the other hand, the strength of the interactions involving charged residues is lowered significantly by a water environment, with the only exception being Arg, whose guanidinium group also has a strong dispersion interaction. The polar and sulfuric groups are shifted toward lower stability, whereas the smaller residues of the same kind are moved more (Asn more than Gln, Ser more than Thr, and Cys more than Met). This can be accounted for by the less extensive dispersion interactions.

Interaction Energy Distributions in the Gas Phase. A major question of this study is how the selected cluster representatives are relevant to the overall energy distribution for all of the interacting residues with a particular amino acid, or better said how representative these interactions are. We have shown previously that the interaction energy of the cluster representative is a reasonable approximation of the interaction energy of the whole selected cluster for one particular pair.¹⁸

The calculation of the 20×20 interaction matrix of the cluster representatives shows that some interactions are not symmetrical in terms of their energies. This is more profound for the polar amino acid side chains, namely, Ser—Thr and Thr—Ser. They differ significantly in their interaction energies for cluster representatives (-7.3 vs -1.7 kcal/mol). While the total number of interactions is the same for both pairs, the clustering algorithm apparently provided two different geometries for the cluster representatives. One can expect a symmetry of the interaction

energy values if the ensemble of structures is large enough to result in the same geometry for both representatives obtained by the cluster-analysis algorithm.

Our aim was to describe the cluster representative in the context of the overall geometry distribution for the selected pair of amino acids appearing in proteins. To see just how representative it is of the whole distribution required computing all interaction energies for a given side chain/side chain distribution and comparing with the energy of the representative conformation. The only way to achieve this in a reasonable time was to use the parm03 force field. We chose tryptophan (Trp) as our side chain 1 and calculated its interactions with all 20 amino acid side chains. Trp was chosen because of the reasonable level of agreement between the *ab initio* and force-field results for the calculations involving Trp. Moreover, the interactions with this residue do not show any repulsive interaction energy values in any of the methods used.

The results for the gas phase are presented in Figures 5 as histograms of the calculated interaction energies with parm03 force field. Most of the histograms have one peak slightly below the zero value. Only the interactions of Trp with negatively charged residues have clear two-peak distributions. Most of the distributions of interaction energies seem to be limited by zero on one side and by the cluster representative value at the other extreme. To confirm the behavior by a higher level of methodology, we recalculated the distribution for 100 randomly selected pairs for every 20 amino acids interacting with Trp by RI-DFT-D for the optimized structures. As follows from our analysis (see Figure 6; Figure 7 shows those with the OPLS-AA force field), both distributions are very similar, and the energy limit at the zero value is clearly more distinct for the histograms of the RI-DFT-D energies.

One important conclusion can be made on the basis of the obtained results. The cluster representative values are mostly extreme cases of the side-chain/side-chain interactions and cannot serve as a measure of the interaction-energy distribution or its typical value. Particularly in the case of side chains involved in hydrogen bonds with the Trp (eg Asp, Glu, Ser, and Thr), the representatives tend to correspond to low-energy conformations. Where interactions are less directional, as in interactions involving hydrophobic contacts, the representative does not necessarily have a low energy conformation. The results are summarized in Table 4. We cannot relate the data obtained

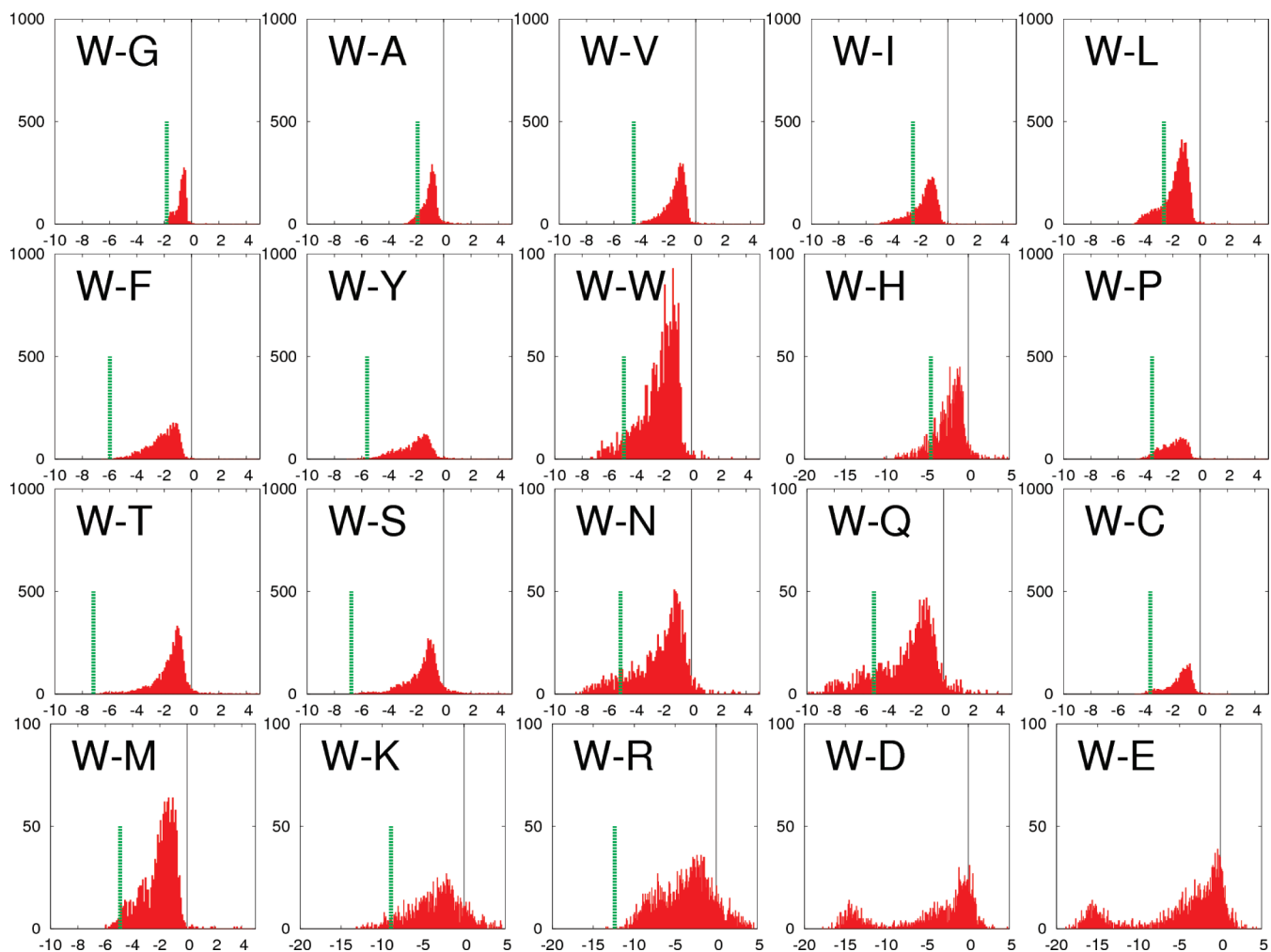


Figure 5. Histograms of the interaction energies of the side chain/side chain interactions of tryptophan with all of the other residues calculated with the parm03 force field in the gas phase. Energies on the x axis are in kcal/mol.

for the representative interactions of a particular amino acid to any of the phenomenological matrices published.^{4,7,11}

Discussion

It is not a simple task to establish properly the meaning of the calculated interaction energies between the amino acid side chains, especially if we take only one particular interaction as the representative. As described earlier, there are two extreme views of the side-chain interactions in proteins. The first extreme is that their arrangement is completely random and mostly the backbone properties dictate the fold; the second view is that these interactions are the basis of intramolecular stabilization of the fold and their positions are energy tuned. On the basis of our results presented here, we see the protein stabilization and fold in proportion to the specific and nonspecific interactions depends on the structural and sequential contexts of the protein in question.

The complete interaction energy matrix for all of the amino acids in proteins supports the view that the cluster representatives describe the important spatial but mostly local interactions selected by the character of the residue to maximize the interaction strength in a well-defined spatial

arrangement. This view is supported by our previous analysis of the cluster representative, which is a maximum of the distribution of the interaction energy for a certain cluster.

Additionally, all of the calculated interaction energies in the matrix were attractive. This is not a trivial finding, even if valid for cluster representatives. The common way of interpreting side-chain/side-chain interactions in proteins is that the resulting interaction is a balance between stabilization and repulsion. Some side chains are displaced in nonfavorable orientations (in extreme cases, they can be repulsive) caused by a much greater influence of the adjacent residues or the secondary structural elements. Our data suggest that this is not the case — at least not for such geometrically exclusive interactions as the calculated set constitutes. A general explanation for protein folding can be attributed to the fact that the sharp character of repulsion does not allow side chains to occupy unfavorable positions and the typical pair geometry in proteins is always adjusted to prevent such an interaction mode.

We are aware of the fact that the benchmark RI-DFT-D values slightly overestimate the interaction energies (by 0.3 kcal/mol on average) for the weakly bound pairs of aliphatic residues such as Ala–Leu as we have proven in a previous paper,¹⁸ and therefore the values are generally higher than

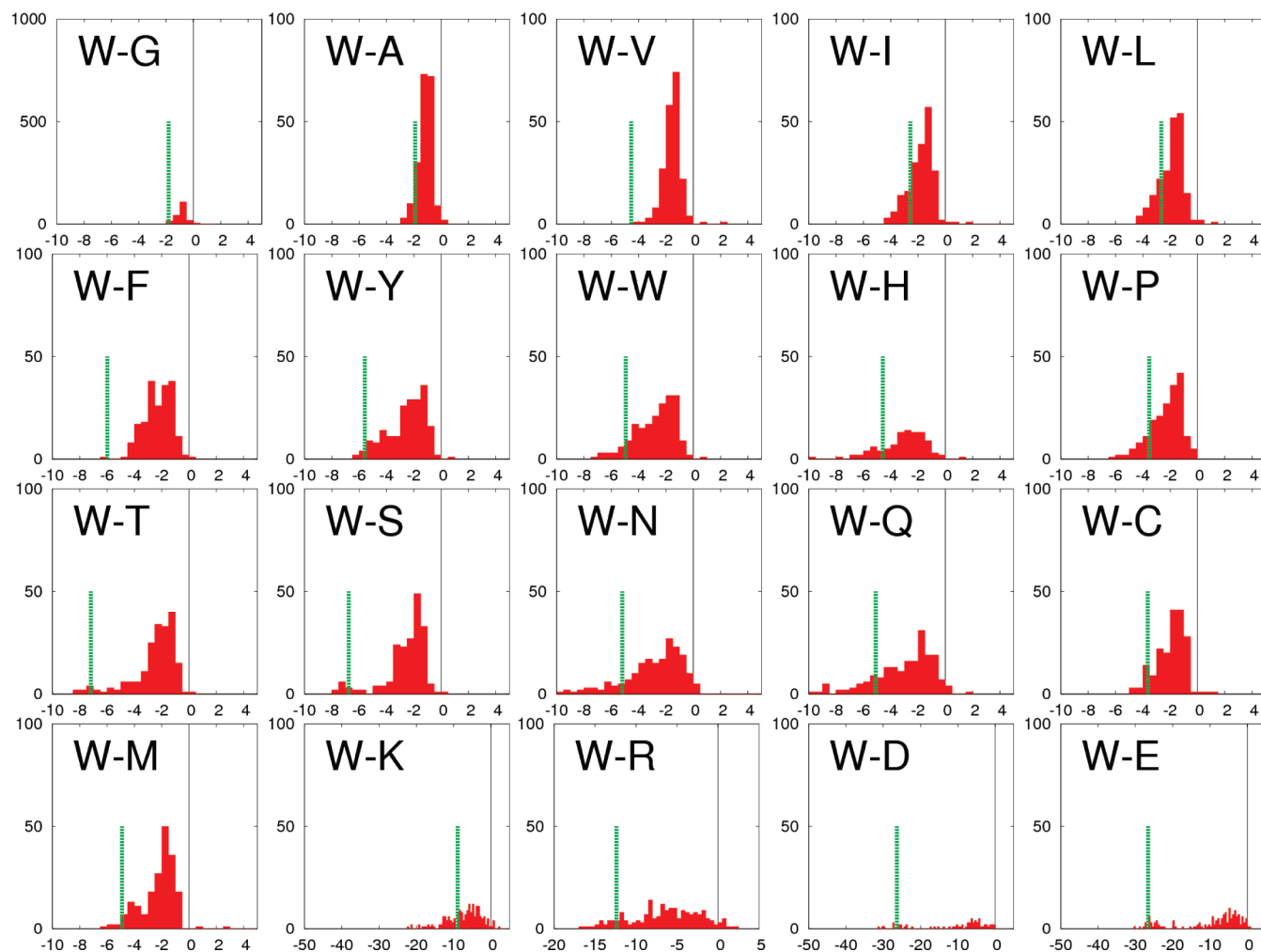


Figure 6. Histograms of the interaction energies of the side chain/side chain interaction of tryptophan with all of the other residues calculated with the RI-DFT-D method in the gas phase. Energies on the x axis are in kcal/mol.

those obtained by empirical potentials. Another fact contributing to this difference is that both force fields are parametrized for a solvent in which the interactions are screened by the environment, namely, the partial charges for the parm03 force field have been parametrized for the dielectric continuum model²⁶ with a dielectric constant equal to 4. Furthermore, the atomic charges for the OPLS-AA/L force field were derived from the parm94 force field. Finally, the non-negligible problem in the utilization of the force field in this study and the interaction energy's accuracy is caused by the addition and optimization of hydrogen atoms. The SCC-DF-TB-D optimization makes them too near the atoms of the other interacting residue. While the optimization of hydrogens had a stabilization effect in the quantum chemical calculations, the opposite is true for the force-field interaction energies.

Although the correlation between the calculated interaction energy matrices is high, especially for the gas phase energies ($r = 0.98$ and 0.95 for OPLS and parm03 in comparison with the RI-DFT-D energies, respectively), the particular differences are not negligible. Fortunately, the interaction energies as a whole are parametrized quite successfully in force fields, but they can vary quite significantly in specific cases.

Unfortunately, some of these specific interactions could be quite important, which is a major problem in the utilization of force fields for the issues addressed by structural biology. One of the reasons for the problematic behavior of the force fields seems to be the repulsive term in the force-field potential form.

Our initial intention was to provide a complete interaction energy matrix for amino acid side-chains and compare it to some extent with the previously published data by Miyazawa and Jernigan and others.^{4,7,11,33,34} This comparison is not possible based solely on the results of our calculations for cluster representatives. We have found that the cluster representatives are not statistically significant for the whole ensemble of interactions. And because we limited our analysis to gas-phase interaction energy as the first approximation, we could only attempt to adjust a significance of the representative values by a calculation of the interaction energy distribution for the complete side-chain interactions of tryptophan.

This comparison, i.e., the interaction energies for the representative geometries and the overall distribution of the interaction energies, showed the significance of cluster representative geometries in the context of the protein and investigated the importance of such interactions. Our results

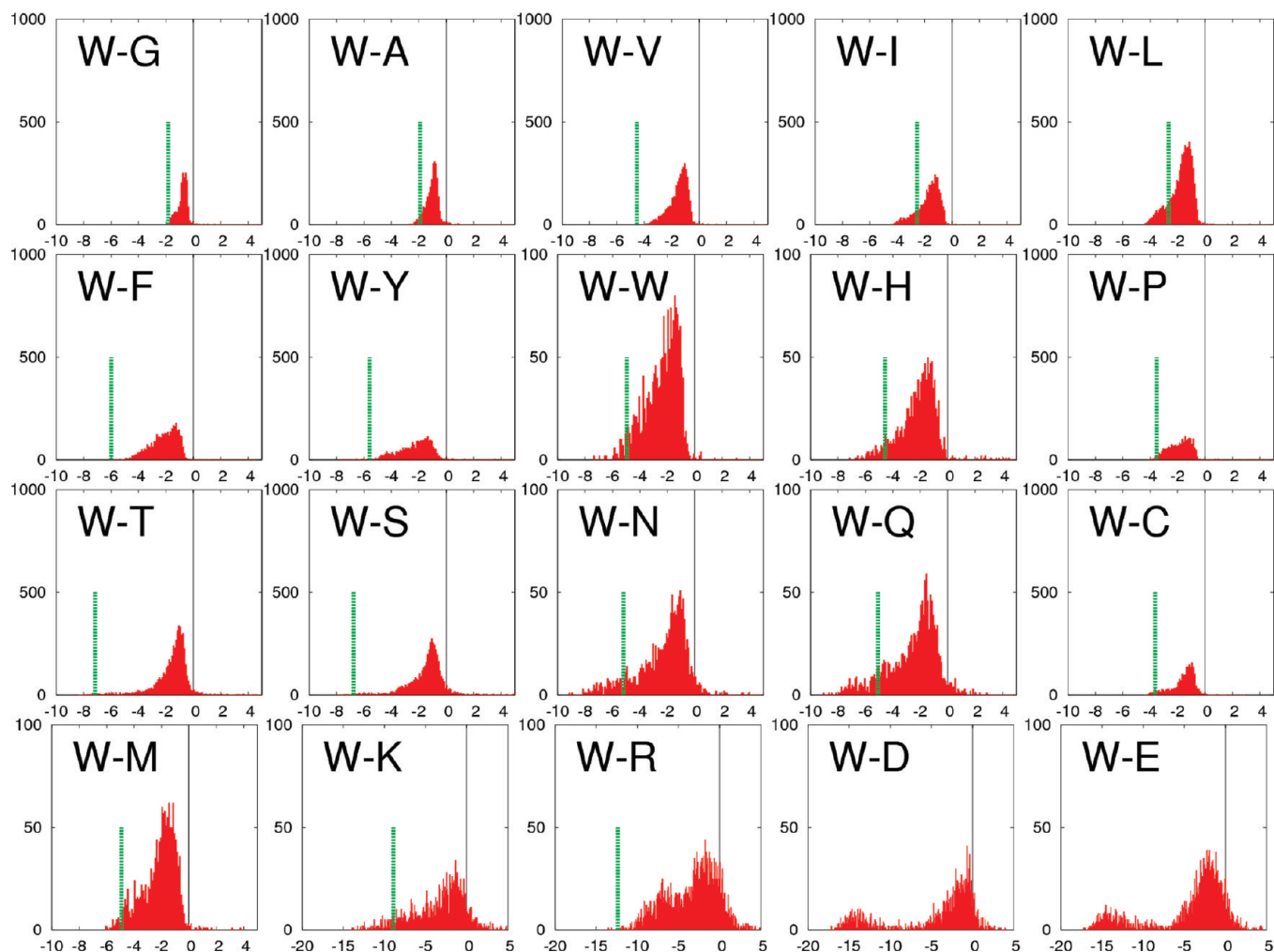


Figure 7. Histograms of all interaction energies of the side chain/side chain interactions of tryptophan with all of the other residues calculated with the OPLS-AA force field in the gas phase. Energies on the x axis are in kcal/mol.

Table 4. Comparison between the Interaction Energies for the Cluster Representatives and the Energetically Most Populated Pairs for All of the Pairs of the Trp Residue

system	WG	WA	WV	WI	WL	WF	WY	WW	WH	WP
IE cluster representative	-1.49	-2.08	-4.01	-2.33	-2.27	-4.72	-4.63	-4.26	-4.37	-2.82
most populated IE	-0.55	-0.85	-1.15	-1.15	-1.35	-1.35	-1.45	-1.35	-2.25	-1.45
system	WT	WS	WN	WQ	WC	WM	WK	WR	WD	WE
IE cluster representative	-6.02	-5.23	-4.78	-5.73	-3.20	-5.39	-5.00	-10.78	-14.86	-17.55
most populated IE	-1.05	-1.15	-1.25	-1.25	-0.75	-1.35	-2.15	-2.35	-1.15	-0.35

led to the conclusion that the optimum-energy side-chain interactions are not the most abundant ones in proteins. They are strong enough to be geometrically as well as energetically distinguishable from the mostly random (and mostly attractive) interactions of the majority of side-chain/side-chain pairs. It is therefore plausible to suggest that the interactions represented by cluster representatives are of crucial importance for protein stability or protein function because of their selectivity and strength.

The distributions of the interaction energies also suggest that the approximations lying behind the phenomenological potentials might simply be wrong, as the distributions are not Boltzmann-like. Therefore, the simple calculation of the free

energies from the detected contacts is not easily connected to the real energies, as has already been indicated by Thomas and Dill.¹³

Conclusions

We have calculated the matrix of interaction energies by means of the RI-DFT-D method as a benchmark and compared it to the same matrix calculated by the parm03 and OPLS-AA/L force fields in the gas phase while utilizing a simple model of different environments. We have further calculated the distributions of the interaction energies for several pairs to discover the meaning of such interactions.

- All of the interaction energies in the gas phase are attractive with the exception of the ones for pairs with the same charges.

- Force fields generally yield good overall interaction energies for the set, but they can have problems in calculations of specific representative interactions.

- Interaction energies are generally lowered by solvent dielectric screening — the lowest difference between the gas phase and environment goes from aliphatics to aromatics and polars, and the biggest difference can be detected for charged residues.

- The histograms of the interaction energies showed that distributions of interaction energies are neither normal nor Boltzmann-like.

- Geometrically chosen cluster representatives are not representatives for the entire side-chain/side-chain interaction distribution. Most probably, they are representatives of the strongest interactions in a protein, often being functionally or structurally important.

Acknowledgment. This work was supported by Grant No. P208/10/0725 from Grant Agency of the Czech Republic and by grant No. LC512 from the Ministry of Education, Youth and Sports (MSMT) of the Czech Republic. It was also a part of the research projects No. Z40550506 and MSM6198959216. It was also part of Institutional Research Concept No. AV0Z505200701 of the Academy of Sciences of the Czech Republic.

Supporting Information Available: Matrices of interactions calculated in OPLS-AA, parm03 force fields and by the PM6-DH method together with Miyazawa-Jernigan contact energies. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (3), 197–208.
- (2) Anfinsen, C. B. Principles That Govern Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230.
- (3) Miyazawa, S.; Jernigan, R. L. A New Substitution Matrix for Protein-Sequence Searches Based on Contact Frequencies in Protein Structures. *Protein Eng.* **1993**, *6* (3), 267–278.
- (4) Miyazawa, S.; Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **1996**, *256* (3), 623–644.
- (5) Miyazawa, S.; Jernigan, R. L. An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins: Struct., Funct., Genet.* **1999**, *36* (3), 357–369.
- (6) Bahar, I.; Kaplan, M.; Jernigan, R. L. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. *Proteins: Struct., Funct., Genet.* **1997**, *29* (3), 292–308.
- (7) Betancourt, M. R. Knowledge-based potential for the polypeptide backbone. *J. Phys. Chem. B* **2008**, *112* (16), 5058–5069.
- (8) Lu, M.; Dousis, A. D.; Ma, J. OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing. *J. Mol. Biol.* **2008**, *376* (1), 288–301.
- (9) Miyazawa, S.; Jernigan, R. L. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins. *J. Chem. Phys.* **2005**, *122* (2), 4012–4030.
- (10) Buchete, N. V.; Straub, J. E.; Thirumalai, D. Orientation-dependent coarse-grained potentials derived by statistical analysis of molecular structural databases. *Polymer* **2004**, *45* (2), 597–608.
- (11) Sippl, M. J. Knowledge-Based Potentials for Proteins. *Curr. Opin. Struct. Biol.* **1995**, *5* (2), 229–235.
- (12) Laskowski, R. A. <http://www.ebi.ac.uk/thornton-srv/databases/sidechains> (accessed January 15, 2009).
- (13) Thomas, P. D.; Dill, K. A. Statistical potentials extracted from protein structures: How accurate are they. *J. Mol. Biol.* **1996**, *257* (2), 457–469.
- (14) Li, D. W.; Bruschweiler, R. A. Dictionary for Protein Side-Chain Entropies from NMR Order Parameters. *J. Am. Chem. Soc.* **2009**, *131* (21), 7226–7232.
- (15) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (18), 6946–6951.
- (16) Morozov, A. V.; Misura, K. M. S.; Tsemekhman, K.; Baker, D. Comparison of quantum mechanics and molecular mechanics dimerization energy landscapes for pairs of ring-containing amino acids in proteins. *J. Phys. Chem. B* **2004**, *108* (24), 8489–8496.
- (17) Singh, J.; Thornton, J. M. *Atlas of Protein Side-Chain Interactions*; IRL Press: Oxford, U. K., 1992.
- (18) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrasek, J. Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated ab Initio Quantum Chemical and Empirical Potential Procedures. *J. Chem. Theory Comput.* **2009**, *5* (4), 982–992.
- (19) van der, S. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.
- (20) Biedermannova, L.; Riley, K. E.; Berka, K.; Hobza, P.; Vondrasek, J. Another role of proline: stabilization interactions in proteins and protein complexes concerning proline and tryptophane. *Phys. Chem. Chem. Phys.* **2008**, *10* (42), 6350–6359.
- (21) Kumar, A.; Elstner, M.; Suhai, S. SCC-DFTB-D study of intercalating carcinogens: Benzo(a)pyrene and its metabolites complexed with the G-C base pair. *Int. J. Quantum Chem.* **2003**, *95* (1), 44–59.
- (22) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a sparse matrix-based implementation of the DFTB method. *J. Phys. Chem. A* **2007**, *111* (26), 5678–5684.
- (23) Cerny, J.; Jurecka, P.; Hobza, P.; Valdes, H. Resolution of identity density functional theory augmented with an empirical dispersion term (RI-DFT-D): a promising tool for studying isolated small peptides. *J. Phys. Chem. A* **2007**, *111* (6), 1146–1154.
- (24) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. Electronic-Structure Calculations on Workstation Computers - the Program System Turbomole. *Chem. Phys. Lett.* **1989**, *162* (3), 165–169.

- (25) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105* (28), 6474–6487.
- (26) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24* (16), 1999–2012.
- (27) Hess, B.; Kutzner, C.; van der, S. D. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.
- (28) Schafer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C. W.; Eckert, F. COSMO Implementation in TURBOMOLE: Extension of an efficient quantum chemical code towards liquid systems. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2187–2193.
- (29) Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z. K.; Gilliland, G.; Weissig, H.; Westbrook, J. The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Biol.* **2000**, *7*, 957–959.
- (30) Kaur, D.; Sharma, P.; Bharatam, P. V. A comparative study on the nature and strength of O-O, S-S, and Se-Se bond. *THEOCHEM* **2007**, *810* (1–3), 31–37.
- (31) Vondrasek, J.; Mason, P. E.; Heyda, J.; Collins, K. D.; Jungwirth, P. The Molecular Origin of Like-Charge Arginine-Arginine Pairing in Water. *J. Phys. Chem. B* **2009**, *113* (27), 9041–9045.
- (32) Berka, K.; Hobza, P.; Vondrasek, J. Analysis of Energy Stabilization inside the Hydrophobic Core of Rubredoxin. *Chemphyschem* **2009**, *10* (3), 543–548.
- (33) Betancourt, M. R. Empirical model of residue contact probabilities for polypeptides. *J. Chem. Phys.* **2010**, *132* (8), 8613–8621.
- (34) Miyazawa, S.; Jernigan, R. L. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins: Struct., Funct., Genet.* **1999**, *34* (1), 49–68.
- (35) Merritt, E. A.; Bacon, D. J. Raster3D: Photorealistic molecular graphics. *Macromol. Crystallogr., Part B* **1997**, *277*, 505–524.
CT100007Y

Interdomain Conformations in the Full-Length MMP-2 Enzyme Explored by Protein–Protein Docking Calculations Using pyDock

Haydee Valdés,[†] Natalia Díaz,[†] Dimas Suárez,^{*,†} and Juan Fernández-Recio[‡]

Dpto. Química Física y Analítica, Universidad de Oviedo, C/Julián Clavería, 8, 33006, Oviedo (Asturias), Spain and Barcelona Supercomputing Center, C/Jordi Girona 29, E-08034 Barcelona, Spain

Received February 18, 2010

Abstract: Current understanding of the collagenolytic activity performed by the matrix metalloproteinases (MMPs) assumes some degree of relative motion between their catalytic and hemopexin-like domains, according to evidence from low-resolution techniques for some of the MMP family members. Herein, we employ protein–protein docking calculations to investigate the structure in aqueous solution of the full-length MMP-2 enzyme in its active form, for which there is not yet experimental evidence of interdomain movement. After docking the domains as free rigid-body subunits, the linker region connecting the catalytic and hemopexin-like domains is taken into account *a posteriori* by merely adding an empiric energy term computed from expected end-to-end distance to the scoring function. Finally, full-length MMP-2 structures are generated by model building the linker residues in the most stable docking poses. The results add support to the hypothesis that the interdomain dynamics of a single MMP-2 molecule in aqueous solution can result in a manifold of conformations, with some preferred orientations. Globally, this structural information could be helpful in future experimental or computational studies aimed to elucidate the dynamical behavior of the MMP-2 enzyme in solution.

Introduction

Matrix metalloproteinases (MMPs) are an important family of zinc- and calcium-dependent peptidases involved in the proteolytic processing of the pericellular environment. The MMPs can cleave virtually all structural matrix proteins (collagen, aggrecan, laminin, etc.), but they also process adhesion molecules (integrins) and biologically active molecules like growth factors, cytokines, and growth factor receptors, contributing thus to the regulation of cellular behavior.^{1,2} Accordingly, they play a central role in different physiological processes, and their expression is also known to increase in various inflammatory, malignant, and degenerative diseases.^{3,4}

All of the MMPs share a significant sequence homology and, in most cases, a common multidomain structure formed by an N-terminal prodomain, a catalytic domain, and a C-terminal hemopexin-like domain joined to the catalytic domain through a linker region (**LK**) of variable length (14–68 residues).^{5,6} The N-terminal pro-peptide blocks the access to the active site cleft in the catalytic domain and is removed upon activation. The catalytic domain (**CAT**, about 170 residues), which holds the proteolytic activity, displays a twisted five-stranded β sheet, three α helices, and several bridging loops. In the gelatinases (MMP-2 and MMP-9), this domain has an additional 175 amino acid residue insert comprising three fibronectin-related type II modules (**FIB1–3**) conferring gelatin and collagen binding properties (see Figure S1 in the Supporting Information). The C-terminal hemopexin-like (**HPX**) domain, which is important in regulating the MMP activation, localization, and inhibition, presents a four-bladed propeller structure. The MMPs also contain a catalytic zinc ion (Zn1), a second zinc ion (Zn2),

* Author to whom correspondence should be addressed. Fax: 34-985-103125. E-mail: dimas@uniovi.es.

[†] Universidad de Oviedo.

[‡] Barcelona Supercomputing Center.

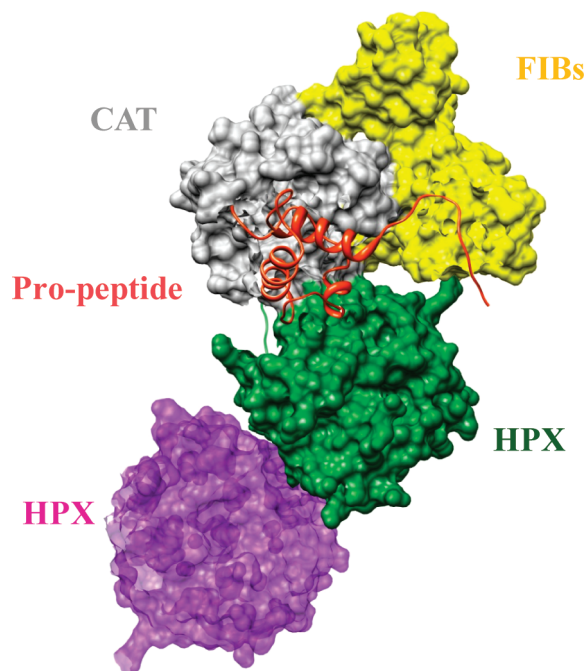


Figure 1. Intermolecular interactions between **HPX** domains of two different molecules in the 1CK7 structure. The second molecule (purple) is only partially represented by the **HPX** domain.

and a number of calcium ions that, according to previous studies, play a structural role in stabilizing several loop regions and fine-tuning the access to the binding site cleft.⁷

To date, X-ray structures of the full-length enzymes have been only reported for MMP-1,^{8,9} MMP-2,^{10,11} and MMP-12.¹² The structures of MMP-1 and MMP-2 display a similar compact arrangement of the catalytic and hemopexin domains, which may suggest the presence of stable interdomain contacts. In contrast, the solid state structure of MMP-12 displays the two domains in a different orientation characterized by a less compact arrangement and a smaller interdomain contact area. Interestingly, some degree of relative motion between the **CAT** and the **HPX** domains of the MMPs has been invoked to explain the collagenolytic activity performed by these enzymes.¹³ This hypothesis has been experimentally confirmed for MMP-1, MMP-9, and MMP-12. Thus, small-angle X-ray scattering and atomic force microscopy experiments have shown that significant interdomain motions occur for MMP-9 in solution.¹⁴ In addition, nuclear magnetic resonance measurements performed for the MMP-1 and MMP-12 enzymes have also revealed that the **CAT** and **HPX** domains experience conformational freedom with respect to each other on the nanosecond time scale.^{12,15}

For MMP-2, an important enzyme in angiogenesis, there is no experimental evidence of interdomain movement yet. The tridimensional structures currently available correspond to the latent pro-enzyme (PDB codes: 1CK7 and 1GXD), where a number of contacts have been observed between the different domains.¹⁰ Thus, the propeptide contacts the **CAT** and **FIB3** domains while remaining close to the first propeller blade of **HPX** (see Figure 1). A larger patch of molecular surface, which amounts to 310 Å² in terms of the reduction of the solvent excluded molecular surface of the

two domains, is covered by the packing of 12 linker residues against the Ω loop of the catalytic domain, resulting in several interdomain H bonds (e.g., Gln₄₃₅–NH⋯Pro₄₆₃–O, Asp₄₁₆–Oδ⋯Thr₄₆₅–OγH) and hydrophobic interactions (e.g., Pro₄₁₇⋯Thr₄₆₅). Finally, the smallest interdomain contact region is formed between the second blade of the **HPX** domain and the first of the three existing **FIB** domains. The only interaction existing in this particular region is a solvent-exposed salt-bridge (Glu₂₄₃⋯Arg₅₅₀). Similarly, in the 1CK7 structure, the propeptide domain only interacts with **HPX** through a salt bridge (Glu₉₅⋯Arg₄₉₅). In addition to these interdomain contacts, which are present in every single MMP-2 molecule in the crystal state, the unit cell of the X-ray structure contains six MMP-2 molecules among which several *intermolecular* contacts between the **HPX** domains of different MMP-2 molecules exist. Interestingly, the most important intermolecular interaction involves the hydrophobic clustering of residues situated at the fourth blade propeller of two different **HPX** domains (Tyr₆₃₆, Leu₆₃₈, Val₆₄₈, Phe₆₅₀, ...), covering 453 Å² of solvent-excluded molecular surface, which is clearly larger than the **CAT**–**HPX** interdomain contact (see Figure 1). In other words, the **HPX** domain contains different areas (patches), some of which are involved in intramolecular interactions and others in intermolecular protein–protein interactions. On the other hand, we have found in a previous computational work that the **CAT**–**HPX** interdomain contact observed in the initial X-ray structure is lost after a 100 ns MD simulation of a fully hydrated MMP-2 molecule, resulting in a quite remarkable rearrangement of the **HPX** domain with respect to the **CAT** and **FIB** domains that adopt an extended conformation during the simulation.¹⁶

Assuming that relatively ample interdomain motions can occur in solution and, as seen above, the **HPX** domain exhibits different regions favorable for protein–protein interactions, it would not be misconceived to think that along with the X-ray structure other conformations could be accessible for a single MMP-2 molecule in solution. Thus, our aim in the present work has been to explore the conformational landscape of the full-length MMP-2 in its active form (i.e., without the pro-peptide), aiming to find feasible conformations that are an alternative to the solid state structure reported experimentally. Clearly, an intensive conformational search using unbiased MD simulations in explicit solvent is computationally too expensive in the case of multidomain proteins like MMP-2, and therefore, other computational techniques should be considered. In this sense, there are several reported computational protein–protein docking methods that are able to efficiently sample alternative orientations between interacting proteins, with the goal of predicting the binding mode of the association.¹⁷ One of the most successful rigid-body docking and scoring schemes, pyDock,¹⁸ has been recently adapted to predict the binding mode of two domains joined by a flexible linker with the addition of end-to-end linker distance restraints, implemented in the module pyDockTET.¹⁹ As an example, the structure of a two-domain protein was proposed to be predicted from homologues of each individual domain in the blind test CAPRI (target 35; <http://www.ebi.ac.uk/msd-srv/capri/>), and

the only successful prediction among all participants was generated by pyDockTET.

Hence, we have used the docking program pyDock, with the module pyDockTET¹⁹ specifically developed to study domain–domain interactions, which has been customized for the present study in order to obtain new models of the MMP-2 enzyme in which the **HPX** domain explores alternative contact regions with the **CAT** and **FIB** domains.

Methods and Computational Details

pyDock Methodology. The first stage of our computational study applied the pyDock docking protocol,¹⁸ which is written in python and uses the MMTK set of python libraries²⁰ for parsing PDB files, calculating united atom AMBER 94 charges, and for other geometry manipulation tasks.

The pyDock docking protocol consists of four steps. In the first one, all the PDB files containing the macromolecules to be docked are preprocessed. This means that only the ATOM records of the 20 standard residues are kept, whereas everything else is removed. In other words, cofactors, hydrogens, and OXT along with HET records are systematically erased. These output PDB files become then the input files for the FTDock²¹ algorithm executed in the second step. FTDock is an algorithm within the 3D-Dock suite of programs designed to enable computational prediction of protein–protein conformations. The FTDock algorithm is based on that of Katchalski-Katzir et al.²² It discretizes the two molecules onto orthogonal grids and performs a global scan of translational and rotational space of possible positions of the two molecules, limited by surface complementarity and an electrostatic filter (optional). The latter is mainly used to discriminate, according to electrostatic favorability, between those complexes (poses) that have similar surface complementarity. FTDock 2.0 was used here, including the electrostatics filter to generate a total of 10 000 rigid-body docking orientations. Those poses are stored in terms of translational coordinates (x , y , z), expressed as integer grid cell displacements of the mobile molecule's center from the center of the static molecule, and rotational angles (z_{twist} , θ , Φ) expressed in degrees. In the third step, each individual geometry of the previously generated 10 000 poses undergoes a coordinate transformation into a suitable format for their use in the fourth and final step (i.e., a rotation and translation matrix for each pose is generated). Here, protein–protein docking poses are scored in order to predict their preferred binding geometry according to the following equation:

$$E = E_{\text{ele}} + E_{\text{desol}} + E_{\text{vdw}} \quad (1)$$

The first term of eq 1 stands for the Coulombic electrostatics where the distance-dependent dielectric constant ($\epsilon = 4r_{ij}$) has been explicitly calculated for all intermolecular atom pairs, with q atomic charges from the AMBER 94²³ force field in elementary charge units and pairwise interaction energy values truncated to a maximum and minimum of +1.0 and –1.0 kcal/mol, respectively, in order to avoid errors from incorrect geometries from the rigid-body approach. E_{desol} represents the effective water-to-interface desolvation

energy,^{24,25} and E_{vdw} is the Lennard-Jones van der Waals energy, also limited to a maximum of +1.0 kcal/mol to allow some interatomic clashes. Typically, E_{vdw} is weighted by a factor of $\omega_{\text{vdw}} = 0.1$ since the van der Waals term is somehow already implicitly included during the FTDock generation of docking poses.

Rigid-body docking poses of multiple domain proteins can be scored by a pseudoenergy term based on restraints derived from linker end-to-end distances.¹⁹ In this method, named pyDockTET (tethered-docking), the scoring function uses the average end-to-end distance, X_m , for a particular linker length (previously derived from a structural database¹⁹) as a restraint to select the correct docking poses. Then, the X_m value and its corresponding standard deviation (σ) are used to develop a function, E_{linker} , which is further incorporated (just by summing it) into the pyDock energy function for the final rescoring of domain–domain poses (for more details on the calculation of E_{linker} , see Figure 8 in ref 19). Essentially, pyDockTET evaluates the linker end-to-end distance for each independent pose, compares it with X_m , and introduces the corresponding energetic penalization according to it.

The conformation of the backbone chains in the rigid-body docking solutions is not refined, as the pyDock protocol has been extensively benchmarked for protein–protein interaction predictions both using internal tests and through the blind competition CAPRI, and the results have given top success rates without needing any refinement. While there have been successful attempts for backbone refinement,²⁶ and the PyDock developing team is actually working on the development of new refinement methods, their current protocol does not seem to significantly improve with respect to the rigid-body scoring.

Two additional tools used in the present work are the Optimal Docking Area (ODA) and the normalized interface propensity (NIP) analyses.^{24,25} The former is a method which enables the examination of any protein surface looking for areas with favorable energy change when buried upon protein–protein association. For that purpose, surface patches with optimal desolvation energy are identified. Such desolvation energy is based on atomic solvation parameters, derived from octanol/water transfer experiments,²⁴ adjusted for protein–protein docking. The NIP method analyzes the 100 lowest-energy solutions (higher-energy solutions do not have an impact on these results) to identify the residues that are most often involved in the docking interfaces, and which are probably involved in protein–protein interactions.

Setting up the pyDock Calculations on the MMP-2 System. Using the pyDockTET method, which as commented above has been specifically developed to study multiple domain proteins, implies that the multidomain molecule has to be split into two subdomains that are expected to be rigid. In the case of the MMPs, this rigid-body approximation is supported by the fact that the secondary structure of the **CAT** and **HPX** domains is very similar in the different X-ray structures regardless of their actual interdomain orientation. Thus, the RMS deviations of the backbone atoms of the **CAT** domain in the 1SU3 (MMP-1), 2CLT (MMP-1), and 3BA0 (MMP-12) structures with respect to that in the 2CLT structure (MMP-2) are 0.62, 0.71,

and 0.92 Å, respectively. The architecture of the **HPX** domain is also well conserved in the same set of structures, the corresponding RMSD values being 1.40, 1.34, and 1.34 Å. Therefore, we decided to formally divide the MMP-2 protein into *molecule A*, composed of the **CAT** and **FIB** domains, and *molecule B*, corresponding to the **HPX** domain. These two *molecules* need to be treated independently as if they are two different proteins. At this stage, the linker region was considered implicitly by adding the pyDockTET energy term to the scoring function.

Another modification was carried out before processing the MMP-2 PDB files. As discussed before, the MMP-2 system contains metallic ions that are already disregarded in the first step of the pyDock protocol. Obviously, this introduces a modification in the overall charge of the molecules to be docked and may, in principle, affect the final results. Indeed, the absence of metallic ions leaves free charged residues in the active site region which turn out to behave as “*strong attractors*”. Consequently, the docking solutions can be biased to conformations where the **HPX** domain is mostly interacting with the active site of the **CAT** domain (see Figure S2 in the Supporting Information) but that do not correspond to the active form of the MMP-2 enzyme. Then, we decided to modify the charge of some of those critical residues in the **CAT** domain in order to overcome this inconvenient. Thus, we mutated, where necessary, glutamic acid into glutamine, histidine into protonated histidine, and aspartic acid into asparagines. In particular, we modified the following residues: (a) Glu₄₀₄ and His₄₁₃, which are coordinated to Zn1; (b) Asp₁₈₀ and His₁₉₃ coordinated to Zn2; (c) Asp₁₈₅ and Glu₂₁₁ coordinated to Ca1; and (d) Glu₁₆₆ and Asp₂₀₄ coordinated to Ca2. Finally, 10 000 rigid-body docking poses were generated by FTDock, and they were further scored according to the pyDock and pyDockTET equations, as above-described.

Building of the MMP-2 Linker. To estimate the influence of the actual linker residues on the relative stability of the most stable poses, as well as to further discriminate among the family of complexes generated by the protein–protein docking calculations, we decided to generate the complete model from the docking solutions by rebuilding the linker region comprising the Asp₄₅₀–Ile₄₆₈ residues that are not taken into account during the pyDock calculations. In addition, bad contacts among the **CAT–FIB** and **HPX** interacting residues were also relaxed. The details of the computational procedure, which are also summarized in Figure S7 in the Supporting Information, are as follows:

1. First, we generated an ensemble of linker structures by means of restrained MD simulations of the following peptide sequence: Ace–Asp–Ile–Asp–Leu–Gly–Thr–Gly–Pro–Thr–Pro–Leu–Gly–Pro–Val–Thr–Pro–Glu–Ile–Nme. The AMBER03 force field,²⁷ which has been used in our previous simulations of the full-length MMP-2 enzyme,¹⁶ was coupled with the Hawkins–Cramer–Truhlar pairwise Generalized-Born (GB) solvent model²⁸ to carry out the MD simulations using the SANDER program included in the AMBER9 suite of programs.²⁹ We defined an end-to-end distance (X_m) as the distance between the C α carbons of the terminal Ace and Nme residues. The value of X_m was

restrained to a specified value using a harmonic biasing potential with a force constant of 10 kcal/mol/Å. We carried out a series of simulation windows beginning at an extended form which corresponds to $X_m = 36.0$ Å. The restrained end-to-end distance was then reduced by 0.25 Å steps down to 5.0 Å (125 windows). The end point of each window was used as a starting point for the next, and each window consisted of 200 ps of equilibration followed by 1.8 ns of production dynamics. The value of the reaction coordinate X_m was saved every 1.0 ps. The biased samplings obtained were used to derive potentials of mean force (PMF) for the end-to-end elongation of the peptide using the Weighted Histogram Analysis method (WHAM).³⁰

2. From each of the most stable pyDock poses (a total of 30 docking solutions), we built a family of full-length MMP-2 structures. To this end, we extracted a set of 40 equally spaced snapshots from the MD simulation of the isolated linker peptide, which were chosen in such way that their reference distance X_m matched the actual C α atoms of the last/first residues of the **CAT/HPX** domains in the corresponding pyDock structures. Each one of the linker structures is connected with the MMP-2 model by superposing the Ace/Nme heavy atoms of the linker peptide onto their counterpart atoms in the terminal **CAT/HPX** residues, and then removing the Ace/Nme coordinates. We also note that in these model building operations, we employed a full atomic representation of the **CAT–FIB** and **HPX** domains including H atoms and metallic ions. At this stage, a total of $30 \times 40 = 1200$ full-length complexes were generated.

3. For every single full-length MMP-2 model, steric clashes between the linker and the **CAT/HPX** atoms, or between the **CAT–FIB** and **HPX** domains, were iteratively identified and relaxed in the following manner. First, the SCWRL4 program³¹ for prediction of protein side-chain conformations was employed to rebuild the side chains of the residues involved in the corresponding steric clash. Then, the coordinates of the same residues were relaxed by carrying out 1000 conjugate gradient steps followed by 25 ps of MD using the AMBER03 force field and a distance dependent dielectric constant ($\epsilon = 4r_{ij}$). A high temperature value (500 K) was used in the restricted MD simulations in order to promote uphill moves of bulky side chains that can be important for properly relaxing some steric collapses. Once the loop over all the steric clashes was completed, the coordinates of the linker atoms and those of the **CAT–HPX–FIB** residues involved in the steric clashes were simultaneously optimized.

4. In principle, the total energy of the partially relaxed full-length MMP-2 models is not useful in obtaining a compensated energetic description because the number and identity of the MMP-2 residues that are structurally relaxed is different in each model. Hence, we combined the interdomain interaction energies (**CAT–FIB**···**HPX** and **CAT–FIB–HPX**···**LK**) and the intrinsic stability of the **LK** region (which is fully relaxed) to assess the stability of the models, defining thus the following scoring function:

$$E = \Delta E_{\text{int}}^{\text{CAT-FIB}\cdots\text{HPX}} + \Delta E_{\text{int}}^{\text{CAT-FIB-HPX}\cdots\text{LK}} + E^{\text{LK}} \quad (2)$$

The required energies were obtained by using the Molecular Mechanical (MM) Poisson–Boltzmann Surface Area (MM-PBSA) approach, which has been applied to perform many classes of approximate binding energy calculations, including protein–protein complexes.³² Hence, we performed single-point MM-PBSA energy calculations using the SANDER program on the whole MMP-2 molecules and on the separated **CAT**–**FIB**, **HPX**, and **LK** domains (capped by Ace/Nme residues; the C- and N-terminal residues of **CAT** and **HPX** are removed). These calculations were performed for all 1200 models using the typical MM-PBSA settings in AMBER9.

Finally, the structural quality information of the most stable full-length MMP-2 models was analyzed by means of the WHAT_CHECK program.³³

Results and Discussion

We applied the pyDock protocol on two different sets of MMP-2 coordinates. On one hand, we used the X-ray (1CK7) structure reported experimentally.¹⁰ For this structure (**XR_MMP-2**), we deleted the coordinates of the propeptide residues, as we want to model the MMP-2 enzyme in its active form (Pro₃₁–Asn₁₀₉). For the same reasons, the N-terminal Tyr₁₁₀ ammonium group was placed interacting with the Asp₄₃₆ as in the so-called “superactivated” form.³⁴ On the other hand, we selected one MD snapshot from our previous MD study (**MD_MMP-2**), which corresponds to an extended configuration accessible for the full-length MMP-2 enzyme in solution.¹⁶ By using two different sets of coordinates, we can assess the sensitivity of the pyDock poses with respect to minor changes in the placement of the residue side chains and in the secondary structure of the protein domains.

For the two structures, **XR_MMP-2** and **MD_MMP-2**, we derived three series of docking solutions differing in the value of the end-to-end average distance (X_m) that is used in pyDockTET. Thus, the family of docking solutions labeled with the **lnk_10** suffix correspond to $X_m \sim 10 \pm 3.2$ Å. Similarly, **lnk_21** and **lnk_25** stand for $X_m \sim 21 \pm 5.0$ Å and $X_m \sim 25 \pm 10$ Å, respectively. These average values and standard deviations were selected from the relation between the length and frequency of linkers in a database of 542 linker structures considered for the parametrization of the pyDockTET treatment. Although the length (number of residues) of the MMP-2 linker is fixed, we note that by using three different end-to-end X_m values, the **HPX** domain is allowed to adopt many more poses beyond those that are compatible with the relatively extended conformation of the linker region in the crystallographic 1CK7 structure ($X_m \sim 36$ Å). For example, the MMP-12 linker, which is only shorter by about three residues than the MMP-2 one, is folded in a more compact arrangement characterized by $X_m \sim 10$ Å.

For each individual combination of X_m distance and initial geometry (e.g., **MD_MMP-2_lnk21**), we analyzed in detail the most stable docking solutions falling within an energy interval of ~ 10 kcal/mol. Typically, 10 different poses fall within such an energy interval. Figure 2 collects the 10 most stable docking solutions (according to the pyDockTET scoring function) obtained for each individual linker restric-

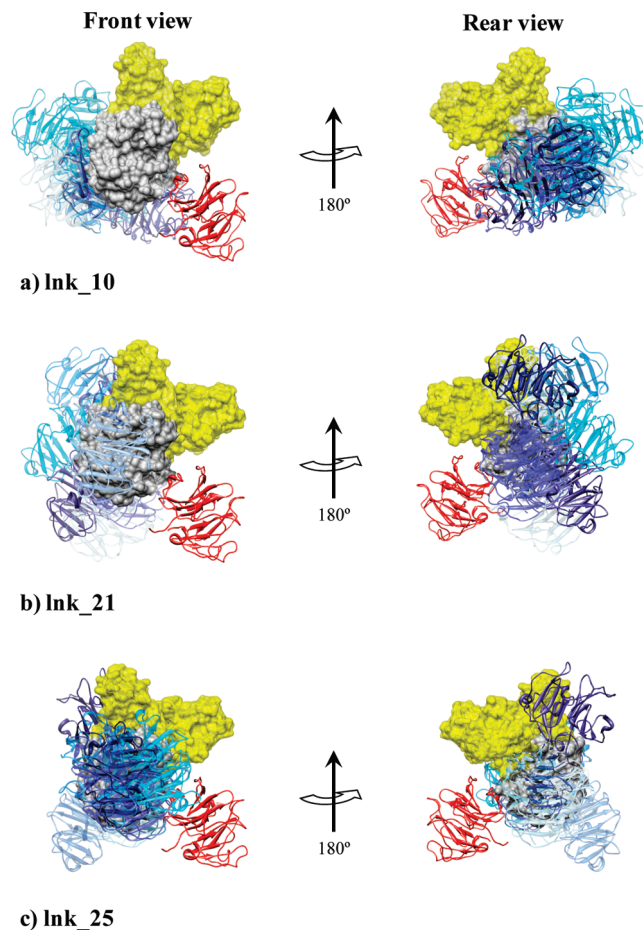


Figure 2. The 10 most stable **HPX** orientations (blue) according to the pyDockTET scoring function (**MD_MMP2**). The position of **HPX** in the X-ray structure is colored in red. Gray and yellow represent the **CAT** and **FIBs** domains, respectively. The gradient of blue colors reflects the different stability (the darker, the most stable) of the various docking solutions. *Front view* refers to the standard orientation of the MMP-2, whereas *rear view* results from rotating the frontal view 180° along the *y* axis.

tion imposed on the **MD_MMP-2** system. Interestingly, quite similar results were obtained for **XR_MMP-2** (see Figure S3 in the Supporting Information), and therefore, we concluded that the pyDock protocol is quite robust and predicts the same kinds of solutions for the MMP-2 system regardless of the actual protein side chain conformations. In what follows, we will present only the results obtained for the **MD_MMP-2** geometry.

Protein–Protein Contacts Favored by the pyDock Calculations. The first observation we can make about the results shown in Figure 2 is that none of the most stable docking solutions predicts an orientation for the **HPX** domain (structures shown in blue) similar to that observed in the crystallographic structure (shown in red). Hence, there is a chance that the **CAT**–**FIB** domains may have another patch suitable for a favorable interaction with **HPX**; i.e., there may exist alternative conformations of the activated and solvated MMP-2 to that reported experimentally for the proenzyme in the solid state. Such affirmation needs, however, a deeper analysis of the data obtained. Inspection of Figure 2 also shows that depending on the formal end-to-end distance of

the linker, the docking solutions explore two different areas of the **CAT-FIB** domains: (a) the active site groove, particularly in case of **lnk_25**, and (b) the rear of the **CAT** and **FIB** domains (mostly for **lnk_10** and **lnk_21**). No docking solutions are placed in the front part of the **FIB** domains.

Solutions where the **HPX** domain is interacting with the active site cleft could be, in a sense, reasonable results provided by the docking calculations given that the MMP-2 active site region actually binds and hydrolyzes other protein systems. Certainly, if the linker adopts an extended conformation with average end-to-end X_m values of 21–25 Å, then the **HPX** domain may be placed in front of the active site (circumstance not possible with a more compact conformation of the linker), but those solutions should be disregarded in our analyses since we are mainly interested in analyzing the active form of the enzyme. Focusing now on the solutions placed in the rear of the **FIB** and **CAT** domains, the following can be observed. In case of large X_m values (**lnk_21** and **lnk_25**), few solutions are placed in a position where the **HPX** is interacting with the rear of the **FIB** domains. More specifically, there are three solutions in the case of **lnk_21** and one solution in the case of **lnk_25**, suggesting that such a region is a potential patch for protein–protein interactions. The remaining solutions are all placed in the rear of the **CAT** domain. This area constitutes then another patch for protein–protein interactions, though it is hard to be precise with a region within the rear of the **CAT** domain where the interaction with **HPX** would be more favorable. Notice, however, that as a general observation we can affirm that the more compact the linker is, the greater the number of structures are interacting with the $\beta 1$ and $\beta 2$ strands of the **CAT** domain. Finally, in terms of abundance, we can establish the following ranking: interactions **HPX–FIB** (H–F) < interactions **HPX–active site** (H–A) < interactions **HPX–CAT** (H–C), where the symbol < implies that a smaller amount of docking solutions show this type of interaction.

As commented upon in the Introduction, the X-ray MMP-12 structure (3BAO) is characterized by a less compact arrangement (in comparison with, e.g., 1CK7) where the **HPX** domain is oriented toward the rear of the **CAT** domain (see Figure S4, Supporting Information). Moreover, in MMP-12, the fourth blade of the **HPX** domain is oriented toward the **CAT** domain as in the case of the most stable docking solutions reported here (see below). Thus, we checked if the orientations of the **HPX** domains of any of the docking solutions collected in Figure 2 matched the orientation of the **HPX** domain in MMP-12. Interestingly, a certain degree of parallelism in the relative orientation of the **HPX** and **CAT** exists between the MMP-12 structure and the MMP-2 docking solutions (see Figure S5, Supporting Information).

Relative Stability of the pyDock Poses. Unfortunately, no systematic behavior exists in terms of energy, meaning that no type of interaction is significantly more stable than any other (see Table 1). For instance, the pose obtained as the most stable solution in the **lnk_25** set, but also that is ranked number eight in **lnk_21**, is of the H–A type. One thing that can be easily seen when analyzing Table 1 is that

Table 1. Relative Energies (kcal/mol) for the 10 Most Stable Docking Solutions in Each of the Three Series of pyDockTET Docking Calculations with Different Linker Restrictions^a

lnk_10	interaction	lnk_21	interaction	lnk_25	interaction
–64.51	H–C	–78.56	H–F	–84.32	H–A
–62.71	H–C	–69.56	H–C	–78.56	H–F
–62.16	H–C	–69.50	H–C	–77.03	H–A
–61.02	H–C	–69.37	H–C	–75.38	H–A
–60.75	H–C	–68.56	H–F	–71.20	H–A
–60.28	H–C	–67.21	H–C	–69.83	H–A
–59.45	H–C	–66.44	H–C	–69.83	H–A
–57.62	H–C	–66.41	H–A	–69.56	H–A
–57.06	H–C	–65.16	H–C	–69.50	H–A
–55.11	H–C	–65.10	H–F	–69.37	H–A

^a Specific domains interacting for each particular solution are indicated in the columns labeled “interaction”.

docking solutions are highly biased depending on the end-to-end distance. For example, the H–C interactions are systematically favored in the **lnk_10** solutions, whereas in the **lnk_25** set the H–A interactions are the most abundant ones. An intermediate situation (**lnk_21**) allows a more even distribution. If for the above commented reasons solutions of the H–A type are disregarded, the most stable solution (H–F type) is found in the **lnk_21** family (this same pose is also the second most stable one in the **lnk_25** ranking). In this structure, the actual value of the linker end-to-end distance measured as the $C\alpha$ – $C\alpha$ distance belonging to the last residue of **CAT–FIB** and the first residue of **HPX** is 25.6 Å, and the **HPX** domain interacts with the rear of the **FIB** domains (see the darkest blue solution in Figure 2, rear view of inset b).

The Fourth Blade in HPX as a Likely Interaction Site. In the most stable pyDock solutions, it turns out that the **HPX** domain interacts preferentially with the first and the second domains of **FIB** via its fourth blade propeller (see Figure S6 in the Supporting Information). Indeed, we have confirmed that in the rest of the 10 most stable solutions (including not only the ones corresponding to the **lnk_21** but also to the **lnk_10** and **lnk_25** sets), **HPX** interacts again via its fourth blade. Interestingly, this is in agreement with the behavior found in the intermolecular **HPX**⋯**HPX** contacts in the 1CK7 crystal structure (see Figure 1) and in the complex between MMP-2 and the TIMP-2 inhibitor,^{11,35} where the **HPX** domain is also interacting via its fourth blade with TIMP-2. However, the intramolecular **HPX**⋯**CAT** interactions in the 1CK7 structure involve mainly the first **HPX** blade. It may be interesting to note that further statistical analysis performed on 5000 docking poses showed that a large number of solutions in which the **HPX** domain is interacting via its first blade are also obtained, but they are less stable than those involving interactions with the fourth blade.

The **HPX** interaction sites can also be characterized by mapping residue NIP values onto the protein surface. We focused on those residues with NIP values between 1 (residues appear in the interface in all the docking solutions) and 0 (residues are found in the docking interfaces as frequently as randomly expected). For **HPX**, the maximum NIP value is 0.3 and corresponds to the Phe₆₅₀ residue. Then,

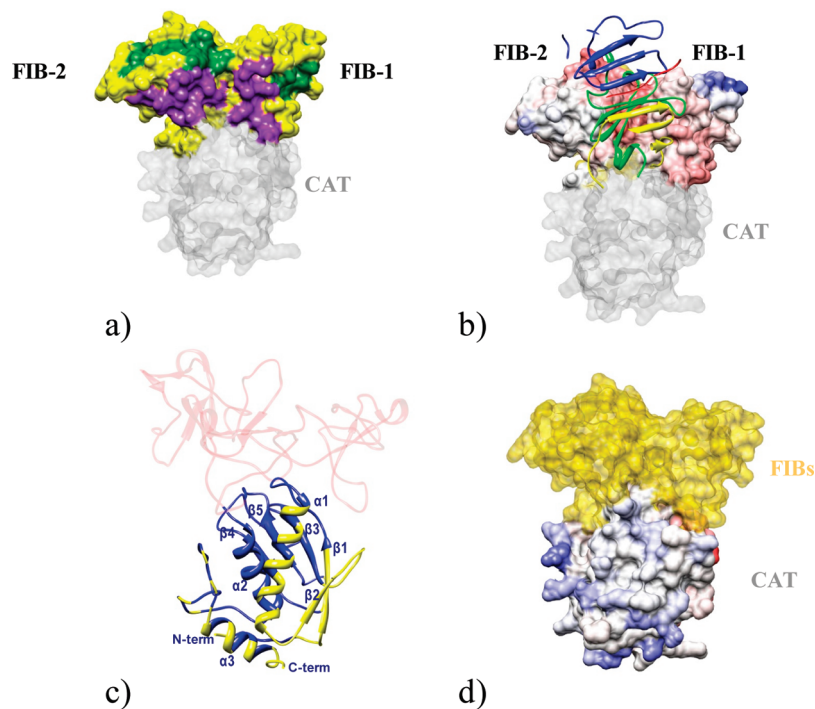


Figure 3. (a) Binding regions of the **FIB** subdomains of MMP-2 colored according to experimental (green) and theoretical (purple) predictions. Gray and yellow surfaces correspond to the **CAT** and the third **FIB** domains, respectively. (b) The **HPX** domain is represented in ribbons and its four blades colored in yellow (first blade), red (second blade), dark blue (third blade), and green (fourth blade). The first and second domains of **FIB** are colored according to ODA desolvation energies. Contrary to the blue areas, the red areas have low desolvation energy. (c) **CAT** domain of MMP-2 colored according to NIP values (yellow means values ranging within 0.4–0.0). (d) ODA representation of the rear of the **CAT** domain.

we checked whether any residue with NIP values ranging between 0 and 0.3 agree with those observed by Piccard et al.³⁵ and Morgunova et al.¹¹ in their experimental study about the interaction of MMP-2 and the TIMP-2 inhibitor. According to Piccard et al., MMP-2 recognizes the so-called GH loop of TIMP-2 by means of Ala₆₁₂, Tyr₆₃₆, Leu₆₃₈, Val₆₄₈, and most significantly, Phe₆₅₀, located at the fourth blade of the **HPX** domain. According to our calculations, those same five residues show the highest NIP values (see Table S1 in the Supporting Information), and more specifically, residue Phe₆₅₀ is playing a major role according to both the experimental and the theoretical studies. Consequently, there is a nice agreement between the theoretical predictions concerning the nature of the **HPX** interaction sites and closely related experimental data.

Protein–Protein Interaction Sites in the CAT–FIB Domains. Few experimental studies have reported information on the gelating binding regions of **FIBs**.^{36,37} Essentially, gelating binding sites have been identified in each of the three **FIB** subdomains and are formed by three conserved clusters of Phe, Trp, and Tyr residues and their surrounding region (colored in green in Figure 3a for **FIB-1** and **FIB-2**). The docking analyses show that **HPX** can interact simultaneously with **FIB-1** and **FIB-2** through a region (colored in purple in Figure 3a) quite close to the gelating binding areas of the two subdomains. For instance, residues Tyr₃₁₄ and Phe₃₃₁ that form part of the hydrophobic cluster of **FIB-2** also present positive NIP values (see Table S1, Supporting Information). Additionally, ODA calculations show that the

region of **FIB** that interacts with **HPX** is a surface patch of optimal desolvation energy (see Figure 3b).

We also looked at the NIP values of the **CAT** domains looking for information that could help to further rationalize the data obtained (see Figure 3c). A few residues belonging to the **CAT** domain show NIP values higher than 0, suggesting, thus, the existence of patches (colored in yellow in Figure 3c) in the **CAT** domain that can likely interact with other proteins. Those patches are quite spread along the $\beta 1$ – $\beta 2$ strands and the $\alpha 1$ and $\alpha 3$ helices (see Figure 3c), with no residues with NIP > 0.4; that is, no “hot spot” residues for protein interactions can be distinguished. This is in consonance with ODA analyses, indicating that the molecular surface of the rear of the **CAT** domain has neutral desolvation energy (see Figure 3d).

Rebuilding the Linker Region. The free energy profile generated with the implicit GB solvent model for the end-to-end elongation of the linker peptide shows only a minimum located at approximately 10.5 Å (see Figure 4). In fact, the PMF profile explores low energy states (<1.25 kcal/mol) within the 10–36 Å range. Although the use of explicit water models would likely improve the accuracy of the conformational free energy differences, the present PMF calculations suggest that the amino acid sequence of the MMP-2 linker region is intrinsically flexible in aqueous solution and that a broad range of end-to-end distances (10–36 Å) could be accessible at room temperature.

Besides estimating the PMF for the end-to-end elongation of the linker peptide, each MD simulation window provided an ensemble of representative structures having the appropri-

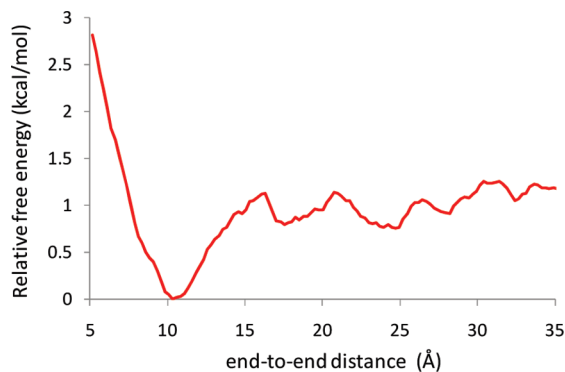


Figure 4. Potential of mean force for the isolated linker peptide.

ate end-to-end distance in order to match the corresponding $C\alpha$ – $C\alpha$ distance between the last **CAT** residue and the first **HPX** residue in the most stable pyDock complexes. In this way, and following the prescriptions detailed in the Methods and Computational Details, we were able to rebuild the full-length MMP-2 enzyme by connecting the **CAT** and **HPX** through a partially relaxed peptide chain as well as fix several bad contacts among side chains in the interaction zone between the **CAT**–**FIB** and **HPX** domains. In the majority of the 1200 full-length MMP-2 structures that were generated from the most stable pyDock poses, interactions between the linker residues and the nearby **CAT/HPX** residues are structurally favorable. However, we also found that several steric clashes could not be fully relaxed through the combination of restrained minimization and MD calculations of the involved residues. Moreover, the backbone of the linker region tends to adopt a strained conformation in many structures.

To translate into a scoring function the diverse structural quality of the full-length MMP-2 models, we estimated both the interaction energies between the **CAT/HPX** domains and the linker region and the intrinsic stability of the linker residues by means of the MM-PBSA method (see Figure 5). However, we note that neither the restricted relaxation of the MMP-2 models generated by the rigid docking calculations nor the limitations of the MM-PBSA methodology for predicting binding or conformational energies of large systems allow us to make clear-cut energetic predictions. Nevertheless, the pyDock structures that are best fitted to accommodate the linker residues should be well captured

by the MM-PBSA scoring function defined in eq 2 thanks to a partial cancellation of errors. Thus, we found that the second solution in the **lnk_10** pyDockTET set, which belongs to the H–C type (see above), generates full-length MMP-2 structures that are much more stable by tenths of a kilocalorie per mole than the other solutions in the same set. The best structure arising from this pyDockTET pose has an end-to-end distance of ~ 13 Å and is characterized by a rather compact arrangement of the linker chain, which is, nevertheless, well fitted to the surrounding protein environment (see Figure 5a). In the case of the **lnk_21** and **lnk_25** sets of pyDockTET solutions, their relatively large end-to-end distances are more compatible with the inclusion of the linker residues, but it turns out that placement of the linker chain stabilizes preferentially the pyDockTET solutions presenting the H–F interaction instead of the H–C one. One of these models with a $C\alpha$ ··· $C\alpha$ separation of 24 Å is shown in Figure 5b, in which we observe how the disposition of the **CAT**–**FIB** and **HPX** domains seems particularly suitable to accommodate the linker chain in a semiextended conformation, which is relatively stable according to the MM-PBSA calculations. Overall, the two full-length MMP-2 models shown in Figure 5 confirm that the interdomain contacts predicted by the pyDockTET solutions are compatible with the actual molecular structure of the linker chain.

Summary and Conclusions

The rigid-body protein–protein docking calculations reported in this work point out that, in the absence of crystallographic contacts and/or other proteins, a single MMP-2 molecule in its active form (i.e., in the absence of the pro-peptide) can adopt different conformations in aqueous solution with respect to that observed by X-ray crystallography. Further details about the actual structure and flexibility of the essential linker region connecting the MMP-2 domains are obtained through a series of model-building operations and MM calculations in which MD snapshots of the linker peptide and the most stable docking solutions are combined. On the basis of our results, we can also draw specific conclusions concerning the multidomain structure of the active form of the MMP-2 enzyme:

- The **HPX** domain tends to interact either with the rear part of the **CAT** domain (preferably with the five-stranded β sheet) or with the first and second **FIB** subdomains.

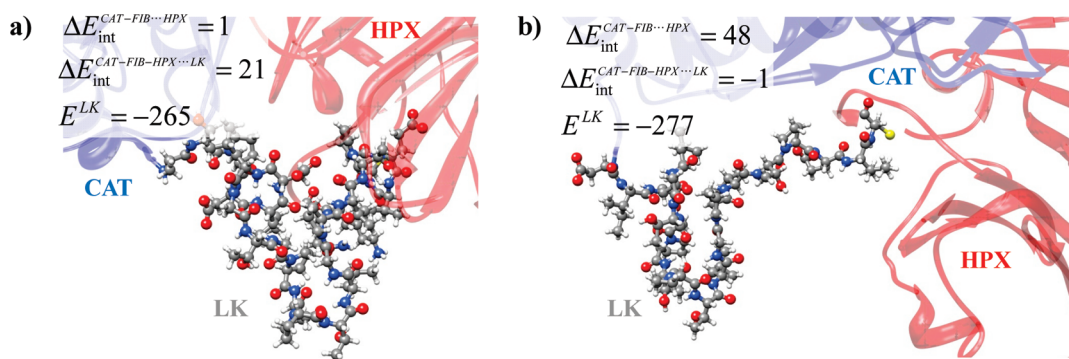


Figure 5. Ball-and-stick views of the linker region adapted to two pyDock solutions corresponding to the **lnk_10** (a) and **lnk_21** (b) sets of **MD_MMP2**. MM-PBSA interdomain interaction and conformational linker energies (kcal/mol) are also indicated.

• The most likely interaction region of the **HPX** domain is constituted by a hydrophobic cluster of residues (Ala₆₁₂, Tyr₆₃₆, Leu₆₃₈, Val₆₄₈, and Phe₆₅₀) located at its fourth blade propeller. In fact, these residues are known to interact either with the TIMP-2 inhibitor or with a second MMP-2 molecule placed in the same crystallographic unit.

• The MD and PMF calculations carried out on the isolated linker region indicate that the linker peptide can easily adopt a large range of end-to-end distances. This seems in consonance with the proposed interdomain flexibility in the MMP-2 enzyme.

• The global interdomain orientations favored by the pyDock calculations are compatible with the molecular structure of the linker residues as confirmed by our full-length MMP-2 models exhibiting a relaxed linker chain connecting their **CAT** and **HPX** domains.

From a methodological point of view, we believe that the computational protocol employed in this work, which is essentially characterized by the rigid-body Pydock calculations and the subsequent reconstruction of linker atoms and removal of bad contacts using all atom MM calculations, could be of interest for other applications. Thus, this strategy, which mixes diversity and likeness in the predicted interdomain conformations, can easily generate a pool of structures for which small-angle X-ray scattering patterns and/or NMR properties could be calculated and used for data analyses of experimental measurements.¹⁵ Similarly, the most-likely Pydock structures could be very useful for further computational studies aimed at the elucidation of the role played by water molecules and protein dynamics in the stability of interdomain conformations. In this respect, we note that the generation of reliable docking structures like those reported in this work for the MMP-2 enzyme could be seen as a prerequisite before carrying out extensive MD simulations in explicit solvent and more sophisticated free energy calculations. As a matter of fact, further computational and experimental work will be required in order to understand the specific roles played by each of the MMP-2 domains during collagenolysis.

Acknowledgment. This research was supported by Grants CTQ2007-63266 and BIO2008-02882 (MICINN, Spain). The authors thankfully acknowledge the computer resources, technical expertise, and assistance provided by the Barcelona Supercomputing Center—Centro Nacional de Supercomputación. H.V. is grateful to Carles Pons (INB—BSC) for his valuable help in using the pyDock program and also thanks MEC (Spain) for her contract and BSC for a mobility program fellowship (BCV-2009-1-0013).

Supporting Information Available: Table S1 and Figures S1–S7. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Sternlicht, M. D.; Werb, Z. *Ann. Rev. Cell. Dev. Biol.* **2001**, *17*, 463–516.
- (2) McCawley, L. J.; Matrisian, L. M. *Curr. Opin. Cell Biol.* **2001**, *13*, 534–540.
- (3) Coussens, L. M.; Werb, Z. *Chem. Biol.* **1996**, *3*, 895–904.
- (4) van Meurs, J.; van Lent, P.; Holthuysen, A.; Lambrou, D.; Bayne, E.; Singer, I.; Berg, W. v. d. *J. Immunol.* **1999**, *163*, 5633–5639.
- (5) Nagase, H.; Woessner, J. F., Jr. *J. Biol. Chem.* **1999**, *274*, 21491–21494.
- (6) Maskos, K. *Biochimie* **2005**, *87*, 249–263.
- (7) Diaz, N.; Suarez, D. *Biochemistry* **2007**, *46*, 8943–8952.
- (8) Jozic, D.; Bourenkov, G.; Lim, N.-H.; Visse, R.; Nagase, H.; Bode, W.; Maskos, K. *J. Biol. Chem.* **2005**, *280*, 9578–9585.
- (9) Iyer, S.; Visse, R.; Nagase, H.; Acharya, K. R. *J. Mol. Biol.* **2006**, *362*, 78–88.
- (10) Morgunova, E.; Tuuttila, A.; Bergmann, U.; Isupov, M.; Lindqvist, Y.; Schneider, G.; Tryggvason, K. *Science* **1999**, *284*, 1667–1670.
- (11) Morgunova, E.; Tuuttila, A.; Bergmann, U.; Tryggvason, K. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7414–7419.
- (12) Bertini, I.; Calderone, V.; Fragai, M.; Jaiswal, R.; Luchinar, C.; Melikian, M.; Mylonas, E.; Svergun, D. *J. Am. Chem. Soc.* **2008**, *130*, 7011–7021.
- (13) Bode, W. *Structure* **1995**, *3*, 527–530.
- (14) Rosenblum, G.; Van den Steen, P. E.; Cohen, S. R.; Grossmann, J. G.; Frenkel, J.; Sertchook, R.; Slack, N.; Strange, R. W.; Opendakker, G.; Sagi, I. *Structure* **2007**, *15*, 1227–1236.
- (15) Bertini, I.; Fragai, M.; Luchinat, C.; Melikian, M.; Mylonas, E.; Sarti, N.; Svergun, D. I. *J. Biol. Chem.* **2009**, *284*, 12821–12828.
- (16) Diaz, N.; Suarez, D.; Valdes, H. *J. Am. Chem. Soc.* **2008**, *130*, 14070–14071.
- (17) Ritchie, D. W. *Curr. Protein Pept. Sci.* **2008**, *9*, 1–15.
- (18) Cheng, T. M.; Blundell, T. L.; Fernandez-Recio, J. *Proteins* **2007**, *68*, 503–515.
- (19) MK Cheng, T.; Blundell, T. L.; Fernandez-Recio, J. *BMC Bioinf.* **2008**, *9*, 441–453.
- (20) Hinsen, K. *J. Comput. Chem.* **2000**, *21*, 79–85.
- (21) Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. *J. Mol. Biol.* **1997**, *272*, 106–120.
- (22) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2195–2199.
- (23) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (24) Fernandez-Recio, J.; Totrov, M.; Abagyan, R. *J. Mol. Biol.* **2004**, *335*, 843–865.
- (25) Fernandez-Recio, J.; Totrov, M.; Skorodumov, C.; Abagyan, R. *Proteins* **2005**, *58*, 134–143.
- (26) Mashich, E.; Nussinov, R.; Wolfson, H. J. *Prot. Struct. Funct. Bioinf.* **2010**, *78*, 1503–1519.
- (27) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A. *J. Comput. Chem.* **2003**, *14*, 1999–2012.
- (28) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P.; Case, P. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.

- (29) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, I. C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER9*; University of California: San Francisco, CA, 2006.
- (30) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, D. R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339–1350.
- (31) Krivov, G. I.; Shapovalov, M. L.; Dunbrack, R. L. *Prot. Struct. Funct. Bioinf.* **2009**, *77*, 778–795.
- (32) Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2003**, *25*, 238–250.
- (33) Hooft, R. W. W.; Vriend, G.; Sander, C.; Abola, E. E. *Nature* **1996**, *381*, 272–272.
- (34) Díaz, N.; Suárez, D. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 50–61.
- (35) Piccard, H.; Van den Steen, P. E.; Opdenakker, G. *J. Leukoc. Biol.* **2007**, *81*, 870–892.
- (36) Gehrmann, M. L.; Douglas, J. T.; Bányai, L.; Tordai, H.; Patthy, L.; Llinás, M. *J. Biol. Chem.* **2004**, *279*, 46921–46929.
- (37) Briknarová, K.; Grishaev, A.; Bányai, L.; Tordai, H.; Patthy, L.; Llinás, M. *Structure* **1999**, *7*, 1235–1245.

CT100097X

New and Efficient Poisson–Boltzmann Solver for Interaction of Multiple Proteins

Eng-Hui Yap* and Teresa Head-Gordon*

Department of Bioengineering, University of California, Berkeley, Berkeley, California 94720, and UCSF and UC Berkeley Joint Graduate Group in Bioengineering

Received March 17, 2010

Abstract: We derive a new numerical approach to solving the linearized Poisson–Boltzmann equation (PBE) by representing the protein surface as a collection of spheres in which the surface charges can then be iteratively solved by new analytical multipole methods previously introduced by us [Lotan, I.; Head-Gordon, T. *J. Chem. Theory Comput.* **2006**, *2*, 541.]. We show that our Poisson–Boltzmann semianalytical method, PB-SAM, realizes better accuracy, more flexible memory management, and at reduced cost relative to either finite difference or boundary element method PBE solvers. We provide two new benchmarks of PBE solution accuracy to test the numerical PBE solutions based on (1) arrays of up to hundreds of spherical low dielectric geometries with asymmetric charges in which mutual polarization is treated exactly and (2) two overlapping spheres with increasing charge asymmetry by solving the PB-SAM method to very high pole order. We illustrate the strength of the PB-SAM approach by computing the potential profile of an array of 60 T1-particle forming monomers of the bromine mosaic virus.

Introduction

The formation of protein complexes is ubiquitous in a crowded, salty cellular environment. Since electrostatic forces dominate the earliest of protein–protein recognition events in the cell, various analytical and numerical continuum theories of bulk electrolytes have been adapted for use to describe protein complexation mechanisms on the supramolecular scale.² One popular continuum mean-field theory is the Poisson–Boltzmann (PB) treatment, which forms the basis of Gouy–Chapman theory^{3,4} in electrochemistry, and under the low field linearized PB (LPB) approximation, the Debye–Hückel theory in solution chemistry⁵ and Derjaguin–Landau–Verwey–Overbeek (DLVO) theory in colloid chemistry.^{6,7} Numerous techniques for solving the PB equation exist,^{1,6} including both analytical or numerical methods, and each has its drawbacks and its strengths.

Analytical methods typically allow rapid solution of the PB equation using multipole expansions under specialized geometries, such as spheres or cylinders. A complete PB

solution comprising one spherical macromolecule was developed by Kirkwood⁸ more than 70 years ago, but generalization of this complete solution to two or more spherical macromolecules proved to be more difficult, and many different partial and approximate solutions have been proposed.^{9–12} We have recently achieved a fundamental result in deriving an analytical PB solution for computing the screened electrostatic interaction between *arbitrary* numbers of spherical proteins of *arbitrarily* complex charge distributions, separated by *arbitrary* distance.¹ While such idealized protein geometries will typically be inappropriate for describing complexation on a supermolecular scale, this new analytical solution is a novel component of our new numerical PB solver for arbitrary protein shape. It also serves as a benchmark for the accuracy of the numerical solutions in certain idealized test cases.

By contrast, numerical methods (see ref 13 for a recent survey), such as finite-difference (FD)^{14–19} and finite-element (FE)^{20–22} methods, can handle arbitrary dielectric boundaries by solving for the PB potential on a 3-D grid or mesh. However, there are limitations of the FE or FD formulations, such as singularities in the potential solution because of point

* To whom correspondence should be addressed. E-mail address: enghui@berkeley.edu (E.-H.Y.); tthead-gordon@lbl.gov (T.H.-G.).

charges, that electric displacement continuity could not be enforced across dielectric boundaries (thereby reducing the solution accuracy and convergence rate), and forces must be estimated from finite-difference calculations.¹³ But most importantly, the requirement that the solution be solved on a grid limits its practical application to spatial domains of either two to three typical macromolecules at reasonably high resolution (~ 0.2 Å) or to larger numbers of macromolecules with greatly diminished resolution and thus solution accuracy. For example, the PBE solution for an assembled 50S ribosomal subunit has been evaluated at 0.45 Å resolution,¹⁴ at the limit of machine memory, but to describe the preceding assembly process that occur over much larger spatial distances, the spatial resolution and consequently the solution accuracy would greatly deteriorate. As such, computational and memory cost in FD and FE methods are strictly functions of the number of grid points and not of the number of macromolecules described.

Boundary element (BE) methods^{22–29} are an attractive alternative since they satisfy both the Dirichlet and von Neuman boundary conditions by construction, singular charges can be correctly treated, and most importantly the 2D solutions on the macromolecular surface removes spatial resolution limitations imposed by the 3D grid of the FD or FE solvers. However increasing the number of boundary surface element results in an increasingly large dense matrix to be solved with severe memory requirements, a problem which scales with the number of macromolecules. Recent acceleration of the BE approach^{24,26} by incorporating fast multipole methods have rendered BE computational times comparable to state-of-the-art software packages like the Adaptive Poisson–Boltzmann Solver (APBS)¹⁴ based on FD solutions.

In this work, we derive a new numerical approach to solving the PB equation by combining the advantages of both the boundary element and our analytical model¹ formalism. In particular, we replace the discretization of the molecule surface into a large number (tens of thousands) of boundary elements, by a discretization involving a smaller number (tens to hundreds) of spheres. The surface charges can then be iteratively solved using analytical multipole methods.¹ We show that our Poisson–Boltzmann semianalytical method, PB-SAM, converges to the analytical solution with better accuracy and at greatly reduced cost relative to the readily available public domain PB solver APBS.¹⁴ Furthermore, we define a high-quality benchmark using 140 poles to describe the electrostatic potential for two overlapping spheres that are models for the sharp features that are sometimes present in real protein geometries, in which we show that our PB-SAM solution converges to the correct solution with the same computational cost or better than the finite difference solution. Finally we illustrate the strength of the PB-SAM approach by computing the potential profile of an array of 60 T1-particle-forming monomers of the bromine mosaic virus (PDB code 1YC6).

Theory

Mathematical Preliminaries. Our theory makes extensive use of the spherical harmonics (SH) family of functions. The

spherical harmonic function of order n and degree m , at polar angle θ and azimuthal angle ϕ , is defined per the convention from Gumerov and Duraiswami³⁰ as

$$Y_{nm}(\theta, \phi) = (-1)^m \sqrt{\frac{(n-|m|)!}{(n+|m|)!}} P_{nm}(\cos \theta) e^{im\phi} \quad (1)$$

where $P_{nm}(x)$ is the *associated Legendre polynomial*. Note that this definition of $Y_{nm}(\theta, \phi)$ differs from the common convention by a $\sqrt{(2n+1)/4\pi}$ factor. The complex conjugate of $Y_{nm}(\theta, \phi)$ will be denoted as $\overline{Y_{nm}(\theta, \phi)}$.

We shall utilize two important properties of spherical harmonics: their addition theorems and orthogonality. Let $\mathbf{r}_1 = [r_1, \theta_1, \phi_1]$ and $\mathbf{r}_2 = [r_2, \theta_2, \phi_2]$ be two points in 3D space specified by spherical coordinates, where $r_2 > r_1$. The Euclidean distance $|\mathbf{r}_1 - \mathbf{r}_2|$ between them then obeys the addition theorems:^{24,31}

$$\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} = \sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{r_1^n}{r_2^{n+1}} \overline{Y_{nm}(\theta_1, \phi_1)} Y_{nm}(\theta_2, \phi_2) \quad (2a)$$

and for the screened Yukawa potential eq 2a is modified to read as

$$\frac{e^{-\kappa|\mathbf{r}_1 - \mathbf{r}_2|}}{|\mathbf{r}_1 - \mathbf{r}_2|} = \sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{r_1^n}{r_2^{n+1}} \hat{i}_n(\kappa r_1) e^{-\kappa r_2} \hat{k}_n(\kappa r_2) \overline{Y_{nm}(\theta_1, \phi_1)} Y_{nm}(\theta_2, \phi_2) \quad (2b)$$

where κ is the inverse Debye–Huckel screening length (described later), and $\hat{k}_n(z)$ and $\hat{i}_n(z)$ are *adapted modified spherical Bessel functions* defined as

$$\hat{k}_n(z) = \sqrt{\frac{2}{\pi}} \frac{e^{-z} z^{n+1/2}}{(2n-1)!!} K_{n+1/2}(z) \quad (3a)$$

$$\hat{i}_n(z) = \sqrt{\frac{\pi}{2}} \frac{(2n+1)!!}{z^{n+1/2}} I_{n+1/2}(z) \quad (3b)$$

$I_n(z)$ and $K_n(z)$ are the *modified Bessel functions of the first and second kind*, respectively. Detailed properties of $\hat{k}_n(z)$ and $\hat{i}_n(z)$ have been described in ref 1.

The spherical harmonic functions are also orthogonal over the surface of a unit sphere (S_1)

$$\int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} Y_{ls}(\theta, \phi) \overline{Y_{nm}(\theta, \phi)} \sin \theta \partial \theta \partial \phi = \frac{4\pi}{2n+1} \delta_{nl} \delta_{ms} \quad (4a)$$

Hence a square-integrable function $g(\theta, \phi)$ on S_1 can be expanded using $\{Y_{nm}\}$ as the basis set

$$g(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{2n+1}{4\pi} G_{nm} Y_{nm}(\theta, \phi) \quad (4b)$$

with the coefficients G_{nm} determined through the reciprocal transform

$$G_{nm} = \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} g(\theta', \phi') \overline{Y_{nm}(\theta', \phi')} \sin \theta' \partial \theta' \partial \phi' \quad (4c)$$

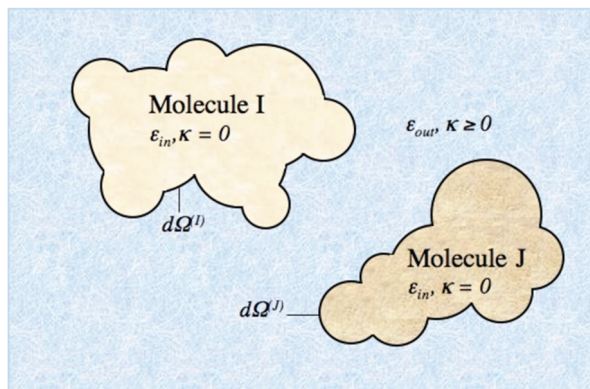


Figure 1. Setting up the boundary value problem. The example system is comprised of two proteins with arbitrary charge distribution, each represented as a collection of overlapping spheres to describe an arbitrarily shaped dielectric boundary containing no salt, immersed in a high dielectric salty continuum solvent. Salt screening effects are captured via the Debye–Huckel parameter κ .

Setting up the Boundary Value Problem. We seek to set up a boundary value problem for a system of N_{mol} macromolecules immersed in an implicit aqueous salty solvent. Figure 1 gives an example of the spatial domain for which we solve the linearized PB equation (LPBE). Each macromolecule I is embedded with $N_C^{(I)}$ fixed partial charge and represented as a collection of $N_S^{(I)}$ overlapping spheres with dielectric constant ϵ_{in} . For simplicity, we consider in this paper the same ϵ_{in} for all molecules, but the model can handle different dielectric constants. The solvent is treated as a continuum with dielectric constant ϵ_{out} , with screening effects because of mobile ions captured via the inverse Debye length κ . The LPBE gives the potential Φ at any point \mathbf{r} in space \mathcal{R}^3 as

$$-\nabla[\epsilon(\mathbf{r})\nabla\Phi(\mathbf{r})] + \kappa^2\Phi(\mathbf{r}) = 4\pi\rho_{\text{fixed}}(\mathbf{r}) \quad (5)$$

where ϵ is the relative dielectric function, ρ_{fixed} is the charge density from the fixed protein partial charges, and $\kappa = \sqrt{8\pi\bar{n}e^2/\epsilon_{\text{out}}k_{\text{B}}T}$, where \bar{n} is the bulk concentration of monovalent salt in the solution, e is the fundamental electronic charge, k_{B} the Boltzmann constant, and T the absolute temperature. Inside each macromolecule I , the potential $\Phi_{\text{in}}^{(I)}(\mathbf{r})$ satisfies the Poisson equation

$$-\nabla^2\Phi_{\text{in}}^{(I)}(\mathbf{r}) = \rho_{\text{fixed}}^{(I)}(\mathbf{r})/\epsilon_{\text{in}} \quad (6a)$$

while in the region outside all macromolecules, the potential $\Phi_{\text{out}}(\mathbf{r})$ satisfies the Helmholtz equation

$$\nabla^2\Phi_{\text{out}}(\mathbf{r}) - \kappa^2\Phi_{\text{out}}(\mathbf{r}) = 0 \quad (6b)$$

We first express the potential $\Phi_{\text{in}}^{(I)}(\mathbf{r})$ anywhere inside molecule I as the sum of the potentials because of the embedded fixed charges and a single-layer of yet unknown reaction charges $f^{(I)}(\mathbf{r})$ on the surface $d\Omega^{(I)}$.^{23,32}

$$\Phi_{\text{in}}^{(I)}(\mathbf{r}) = \sum_{\alpha=1}^{N_C^{(I)}} \frac{1}{|\mathbf{r} - \mathbf{r}_{\alpha}^{(I)}| \epsilon_{\text{in}}} + \frac{1}{4\pi} \int_{d\Omega^{(I)}} \frac{1}{|\mathbf{r} - \mathbf{r}'|} f^{(I)}(\mathbf{r}') d\mathbf{r}' \quad (7)$$

In our new approach, the surface of molecule I is discretized into $N_S^{(I)}$ spheres. We consider each sphere k of molecule I of radius $a^{(I,k)}$ in turn, and all position vectors and coefficients are defined with the center of sphere k as the origin. We apply the first addition theorem (eq 2a) to eq 7 to obtain

$$\Phi_{\text{in}}^{(I,k)}(\mathbf{r}) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(\frac{E_{nm}^{(I,k)}}{r} \left(\frac{a^{(I,k)}}{r} \right)^n + \left(\frac{r}{a^{(I,k)}} \right)^n \text{LE}_{nm}^{(I,k)} \right) Y_{nm}^{(I,k)}(\theta, \phi) + \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(\left(\frac{r}{a^{(I,k)}} \right)^n (\text{LF}_{nm}^{(I,k)} + \text{LFS}_{nm}^{(I,k)}) \right) Y_{nm}^{(I,k)}(\theta, \phi) \quad (8)$$

with the coefficients defined as

$$E_{nm}^{(I,k)} \equiv \sum_{\alpha=1}^{N_C^{(I,k)}} \frac{q_{\alpha}}{\epsilon_{\text{in}}} \left(\frac{r_{\alpha}}{a^{(I,k)}} \right)^n \overline{Y_{nm}^{(I,k)}}(\theta_{\alpha}, \phi_{\alpha}) \quad (8a)$$

$$\text{LE}_{nm}^{(I,k)} \equiv \sum_{\alpha=1}^{\bar{N}_C^{(I,k)}} \frac{q_{\alpha}}{\epsilon_{\text{in}}} \frac{1}{r_{\alpha}} \left(\frac{a^{(I,k)}}{r_{\alpha}} \right)^n \overline{Y_{nm}^{(I,k)}}(\theta_{\alpha}, \phi_{\alpha}) \quad (8b)$$

$$\text{LF}_{nm}^{(I,k)} \equiv \frac{1}{4\pi} \int_{d\Omega^{(I,k)}} \frac{f^{(I,k)}(\mathbf{r}')}{r'} \left(\frac{a^{(I,k)}}{r'} \right)^n \overline{Y_{nm}^{(I,k)}}(\theta', \phi') d\mathbf{r}' \quad (8c)$$

$$\text{LFS}_{nm}^{(I,k)} \equiv \frac{1}{4\pi} \int_{d\Omega^{(I,k)}} \frac{f^{(I,k)}(\mathbf{r}')}{r'} \left(\frac{a^{(I,k)}}{r'} \right)^n \overline{Y_{nm}^{(I,k)}}(\theta', \phi') d\mathbf{r}' \quad (8d)$$

Notice that we have scaled the terms with r_{α}^n and r_{α}^{n+1} dependence by $(a^{(I,k)})^n$ and $(a^{(I,k)})^{-n}$, respectively. This is to avoid machine imprecision as n becomes large. Coefficients with $(r_{\alpha}/a^{(I,k)})^n$ dependence, such as $E_{nm}^{(I,k)}$, are known as multipole (external) coefficients, while those with $a^{(I,k)n}/r_{\alpha}^{n+1}$ dependence ($\text{LE}_{nm}^{(I,k)}$, $\text{LF}_{nm}^{(I,k)}$, and $\text{LFS}_{nm}^{(I,k)}$) are known as Taylor (local) coefficients. The first sum in eq 8 represents the potential due to fixed charges, where $E_{nm}^{(I,k)}$ sums over $N_C^{(I,k)}$ fixed charges *inside* sphere k of molecule I , while $\text{LE}_{nm}^{(I,k)}$ sums over the remaining $\bar{N}_C^{(I,k)}$ fixed charges *outside* sphere k . The second sum in eq 8 gives the potential resulting from the unknown surface charge $f^{(I)}(\mathbf{r})$; $\text{LFS}_{nm}^{(I,k)}$ and $\text{LF}_{nm}^{(I,k)}$ account for represents reactive charges on sphere k and on other spheres in molecule I , respectively.

In the solvent region outside the molecules, the potential $\Phi_{\text{out}}(\mathbf{r})$ can be represented as the sum of Yukawa potentials because of each molecule's yet unknown effective surface charges $h^{(I)}(\mathbf{r})$.^{23,32}

$$\Phi_{\text{out}}(\mathbf{r}) = \sum_{I=1}^{N_{\text{mol}}} \left(\frac{1}{4\pi} \int_{d\Omega^{(I)}} \frac{e^{-\kappa|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r} - \mathbf{r}'|} h^{(I)}(\mathbf{r}') d\mathbf{r}' \right) \quad (9)$$

The above equation valid for the *exposed* portion of sphere k of molecule I . Applying addition theorem 2 (eq 2b) to eq 9, the potential on the exposed surface can be expressed as

$$\Phi_{\text{out}}^{(I,k)}(\mathbf{r}) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(\frac{H_{nm}^{(I,k)}}{r} \left(\frac{a^{(I,k)}}{r} \right)^n e^{-\kappa r} \hat{k}_n(\kappa r) + \left(\frac{r}{a^{(I,k)}} \right)^n \hat{i}_n(\kappa r) (\text{LH}_{nm}^{(I,k)} + \text{LHN}_{nm}^{(I,k)}) Y_{nm}^{(I,k)}(\theta, \phi) \right) \quad (10)$$

where the coefficients are defined as

$$H_{nm}^{(I,k)} \equiv \frac{1}{4\pi} \int_{\text{d}\Omega^{(I,k)}} h^{(I,k)}(\mathbf{r}') \left(\frac{r'}{a^{(I,k)}} \right)^n \hat{i}_n(\kappa r') \overline{Y_{nm}^{(I,k)}}(\theta', \phi') \text{d}\mathbf{r}' \quad (10a)$$

$$\text{LH}_{nm}^{(I,k)} \equiv \frac{1}{4\pi} \int_{\text{d}\Omega^{(I,k)}} \frac{h^{(I,k)}(\mathbf{r}')}{r'} \left(\frac{a^{(I,k)}}{r'} \right)^n e^{-\kappa r'} \hat{k}_n(\kappa r') \overline{Y_{nm}^{(I,k)}}(\theta', \phi') \text{d}\mathbf{r}' \quad (10b)$$

$$\text{LHN}_{nm}^{(I,k)} \equiv \sum_{J \neq I} \sum_{l=1}^{N_K^{(J)}} \left(\frac{1}{4\pi} \int_{\text{d}\Omega^{(J,l)}} \frac{h^{(J,l)}(\mathbf{r}')}{r'} \times \left(\frac{a^{(I,k)}}{r'} \right)^n e^{-\kappa r'} \hat{k}_n(\kappa r') \overline{Y_{nm}^{(I,k)}}(\theta', \phi') \text{d}\mathbf{r}' \right) \quad (10c)$$

The multipole coefficient $H_{nm}^{(I,k)}$ represents effective polarization charges on sphere k of molecule I 's exposed surface. The local coefficients $\text{LH}_{nm}^{(I,k)}$ and $\text{LHN}_{nm}^{(I,k)}$ represent effective polarization charges on other spheres in molecule I and on other molecules, respectively.

With eqs 8 and 10 in hand, we can impose boundary conditions at the dielectric boundary surface $\mathbf{r}_E = (a^{(I,k)}, \theta_E, \phi_E) \in \text{d}\Omega_E^{(I,k)}$ between each sphere k in molecule I exposed to solvent:

$$\Phi_{\text{in}}^{(I,k)}(\mathbf{r}_E) = \Phi_{\text{out}}^{(I,k)}(\mathbf{r}_E) \quad (11a)$$

$$\varepsilon \frac{\text{d}\Phi_{\text{in}}^{(I,k)}}{\text{d}n} \Big|_{\mathbf{r}_E} = \frac{\text{d}\Phi_{\text{out}}^{(I,k)}}{\text{d}n} \Big|_{\mathbf{r}_E}, \quad \varepsilon = \varepsilon_{\text{in}}/\varepsilon_{\text{out}} \quad (11b)$$

The Dirichlet boundary condition (eq 11a) enforces potential continuity across the boundary

$$\sum_{n=0}^{\infty} \sum_{m=-n}^n (E_{nm}^{(I,k)} + a^{(I,k)}(\text{LE}_{nm}^{(I,k)} + \text{LF}_{nm}^{(I,k)} + \text{LFS}_{nm}^{(I,k)})) Y_{nm}^{(I,k)}(\theta_E, \phi_E) = \sum_{n=0}^{\infty} \sum_{m=-n}^n (H_{nm}^{(I,k)} e^{-\kappa a^{(I,k)}} \hat{k}_n(\kappa a^{(I,k)}) + a^{(I,k)} \hat{i}_n(\kappa a^{(I,k)}) (\text{LH}_{nm}^{(I,k)} + \text{LHN}_{nm}^{(I,k)})) Y_{nm}^{(I,k)}(\theta_E, \phi_E) \quad (12a)$$

while the von Neumann boundary condition (eq 11b) enforces electric displacement continuity

$$\varepsilon \sum_{n=0}^{\infty} \sum_{m=-n}^n (-n+1) E_{nm}^{(I,k)} + n F_{nm}^{(I,k)} + n a^{(I,k)} (\text{LE}_{nm}^{(I,k)} + \text{LF}_{nm}^{(I,k)}) Y_{nm}^{(I,k)}(\theta_E, \phi_E) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(H_{nm}^{(I,k)} e^{-\kappa a^{(I,k)}} [n \hat{k}_n(\kappa a^{(I,k)}) - (2n+1) \hat{k}_{n+1}(\kappa a^{(I,k)})] + a^{(I,k)} \left[n \hat{i}_n(\kappa a^{(I,k)}) + \frac{(\kappa a^{(I,k)})^2 \hat{i}_{n+1}(\kappa a^{(I,k)})}{2n+3} \right] \times (\text{LH}_{nm}^{(I,k)} + \text{LHN}_{nm}^{(I,k)}) \right) Y_{nm}^{(I,k)}(\theta_E, \phi_E) \quad (12b)$$

We have introduced $F_{nm}^{(I,k)} \equiv a^{(I,k)} \text{LFS}_{nm}^{(I,k)}$. We continue to simplify eqs 12a and 12b by rearranging

$$\sum_{n=0}^{\infty} \sum_{m=-n}^n (-H_{nm}^{(I,k)} e^{-\kappa a^{(I,k)}} \hat{k}_n(\kappa a^{(I,k)}) + F_{nm}^{(I,k)} + \text{XH}_{nm}^{(I,k)}) Y_{nm}^{(I,k)}(\theta_E, \phi_E) = 0 \quad (13a)$$

$$\sum_{n=0}^{\infty} \sum_{m=-n}^n (e^{-\kappa a^{(I,k)}} [n \hat{k}_n(\kappa a^{(I,k)}) - (2n+1) \hat{k}_{n+1}(\kappa a^{(I,k)})] H_{nm}^{(I,k)} - n \varepsilon F_{nm}^{(I,k)} + \text{XF}_{nm}^{(I,k)}) Y_{nm}^{(I,k)}(\theta_E, \phi_E) = 0 \quad (13b)$$

where

$$\text{XH}_{nm}^{(I,k)} \equiv E_{nm}^{(I,k)} + a^{(I,k)} (\text{LE}_{nm}^{(I,k)} + \text{LF}_{nm}^{(I,k)} - a^{(I,k)} \hat{i}_n(\kappa a^{(I,k)}) (\text{LH}_{nm}^{(I,k)} + \text{LHN}_{nm}^{(I,k)})) \quad (14a)$$

$$\text{XF}_{nm}^{(I,k)} \equiv a^{(I,k)} \left[n \hat{i}_n(\kappa a^{(I,k)}) + \frac{(\kappa a^{(I,k)})^2 \hat{i}_{n+1}(\kappa a^{(I,k)})}{2n+3} \right] (\text{LH}_{nm}^{(I,k)} + \text{LHN}_{nm}^{(I,k)}) + (n+1) \varepsilon E_{nm}^{(I,k)} - n \varepsilon a^{(I,k)} (\text{LE}_{nm}^{(I,k)} + \text{LF}_{nm}^{(I,k)}) \quad (14b)$$

The boundary equations above are valid on the solvent-exposed surfaces of sphere k on molecule I . We need another set of boundary equations on the buried surface $\mathbf{r}_B = [a^{(I,k)}, \theta_B, \phi_B] \in \text{d}\Omega_B^{(I,k)}$. We shall utilize the fact that there is no polarization charge on the buried surface, that is, $f^{(I,k)}(\mathbf{r}_B) = h^{(I,k)}(\mathbf{r}_B) = 0$, since there is no dielectric discontinuity. It follows that scaled versions of the charge distributions, $\tilde{f}^{(I,k)}(\theta, \phi) \equiv (a^{(I,k)})^2 f^{(I,k)}(a^{(I,k)}, \theta, \phi)$ and $\tilde{h}^{(I,k)}(\theta, \phi) \equiv (a^{(I,k)})^2 h^{(I,k)}(a^{(I,k)}, \theta, \phi)$, are also zero on the buried surface. Separately, we can express $\tilde{f}^{(I,k)}$ and $\tilde{h}^{(I,k)}$ in terms of $F_{nm}^{(I,k)}$ and $H_{nm}^{(I,k)}$ using eqs 4c, 8d, and 10a

$$\tilde{f}^{(I,k)}(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{2n+1}{4\pi} F_{nm}^{(I,k)} Y_{nm}^{(I,k)}(\theta, \phi) \quad (15a)$$

$$\tilde{h}^{(I,k)}(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{2n+1}{4\pi} \frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})} Y_{nm}^{(I,k)}(\theta, \phi) \quad (15b)$$

so the “zero-charge” requirement at the buried boundary can be imposed as

$$\sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{2n+1}{4\pi} F_{nm}^{(I,k)} Y_{nm}^{(I,k)}(\theta_B, \phi_B) = 0 \quad (16a)$$

$$\sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{2n+1}{4\pi} \frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})} Y_{nm}^{(I,k)}(\theta_B, \phi_B) = 0 \quad (16b)$$

Equations 13a, 13b, 16a, and 16b specified the complete boundary value problem, from which $F_{nm}^{(I,k)}$ and $H_{nm}^{(I,k)}$ can be solved.

Solution of the Boundary Value Coefficients and Interaction Energy. To solve for $F_{nm}^{(I,k)}$ and $H_{nm}^{(I,k)}$, we need to cast the boundary value problem as a linear system of equations. The infinite expansion series must first be truncated at a maximum pole order p , chosen depending on the desired level of accuracy versus computational cost (see

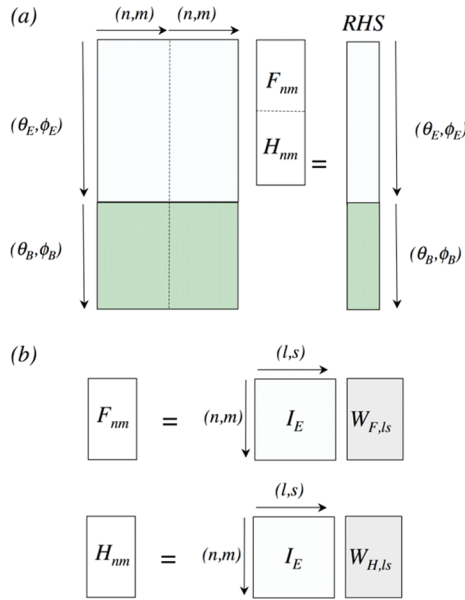


Figure 2. Setting up the boundary equation (eqs 13a, 13b, 16a, and 16b): (a) as a linear least square solve problem and (b) as a matrix-vector multiply operation.

Results). The obvious approach is to set up the boundary equations as a linear least-squares problem (Figure 2a), by discretizing sphere k into M_B buried and M_E exposed grid points, and then finding solutions of vectors $\mathbf{F}_{(l,k)}$ and $\mathbf{H}_{(l,k)}$ that best satisfy the appropriate boundary equations on all grid points. Using the DGELSY routine (complete orthogonal factorization) in LAPACK for $(M_E + M_B) = 10000$ and $p = 60$, each sphere is solved in approximately 10 min. This is computationally intractable if the LPBE needs to be solved repeatedly for tens to hundreds of spheres during dynamics simulations.

Instead, we formulated a novel approach that makes use of spherical harmonics' orthogonal property (eq 4). It converts the problem to a direct matrix-vector multiply operation (Figure 2b), which can be evaluated two-orders of magnitude faster than the LLS approach. We first add $\sum_{n=0}^{\infty} \sum_{m=-n}^n (2n+1) (H_{nm}^{(l,k)}) / (\hat{i}_n(\kappa a^{(l,k)})) Y_{nm}^{(l,k)}(\theta_E, \phi_E)$ to both sides of eq 13a and divide by 4π to arrive at

$$\sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{2n+1}{4\pi} \frac{H_{nm}^{(l,k)}}{\hat{i}_n(\kappa a_{ki}^{(l,k)})} Y_{nm}^{(l,k)}(\theta_E, \phi_E) = \tilde{w}_{H,\text{exposed}}(\theta_E, \phi_E) \quad (17a)$$

where

$$\tilde{w}_{H,\text{exposed}}(\theta, \phi) = \frac{1}{4\pi} \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(H_{nm}^{(l,k)} \left(\frac{2n+1}{\hat{i}_n(\kappa a^{(l,k)})} - e^{-\kappa a^{(l,k)}} \hat{k}_n(\kappa a^{(l,k)}) \right) + F_{nm}^{(l,k)} + XH_{nm}^{(l,k)} \right) Y_{nm}^{(l,k)}(\theta, \phi) \quad (17b)$$

Similarly, we add $\sum_{n=0}^{\infty} \sum_{m=-n}^n (2n+1) F_{nm}^{(l,k)} Y_{nm}^{(l,k)}(\theta_E, \phi_E)$ to both sides of eq 13b and then divide by 4π

$$\sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{2n+1}{4\pi} F_{nm}^{(l,k)} Y_{nm}^{(l,k)}(\theta_E, \phi_E) = \tilde{w}_{F,\text{exposed}}(\theta_E, \phi_E) \quad (18a)$$

$$\tilde{w}_{F,\text{exposed}}(\theta, \phi) = \frac{1}{4\pi} \sum_{n=0}^{\infty} \sum_{m=-n}^n \left(e^{-\kappa a^{(l,k)}} [n \hat{k}_n(\kappa a^{(l,k)}) - (2n+1) \hat{k}_{n+1}(\kappa a^{(l,k)})] H_{nm}^{(l,k)} + (2n+1 - n\epsilon) F_{nm}^{(l,k)} + X F_{nm}^{(l,k)} \right) Y_{nm}^{(l,k)}(\theta, \phi) \quad (18b)$$

Equations 17a and 17b (and similarly 18a and 18b) now completely describe functions $\tilde{w}_H(\theta, \phi)$ (and $\tilde{w}_F(\theta, \phi)$) over the entire surface of sphere k

$$\sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{2n+1}{4\pi} \left[\frac{H_{nm}^{(l,k)}}{\hat{i}_n(\kappa a^{(l,k)})} \right] Y_{nm}^{(l,k)}(\theta, \phi) = \tilde{w}_H(\theta, \phi) = \begin{cases} \tilde{w}_{H,\text{exposed}}(\theta, \phi), & (\theta, \phi) \in \{\theta_E, \phi_E\} \\ 0, & (\theta, \phi) \in \{\theta_B, \phi_B\} \end{cases} \quad (19a)$$

$$\sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{2n+1}{4\pi} [F_{nm}^{(l,k)}] Y_{nm}^{(l,k)}(\theta, \phi) = \tilde{w}_F(\theta, \phi) = \begin{cases} \tilde{w}_{F,\text{exposed}}(\theta, \phi), & (\theta, \phi) \in \{\theta_E, \phi_E\} \\ 0, & (\theta, \phi) \in \{\theta_B, \phi_B\} \end{cases} \quad (19b)$$

The above equations now have the familiar form of spherical harmonic expansion of eq 4b, so we can directly evaluate the coefficients in square parentheses via the reciprocal transform eq 4c. We show below the derivation for $\mathbf{H}^{(l,k)}$

$$\begin{aligned} \frac{H_{nm}^{(l,k)}}{\hat{i}_n(\kappa a^{(l,k)})} &= \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} \tilde{w}_H(\theta', \phi') \overline{Y_{nm}^{(l,k)}}(\theta', \phi') \sin \theta' d\theta' d\phi' = \\ &= \int_{\phi_E} \int_{\theta_E} \tilde{w}_{H,\text{exposed}}(\theta', \phi') \overline{Y_{nm}^{(l,k)}}(\theta', \phi') \sin \theta' d\theta' d\phi' \\ &= \int_{\phi_E} \int_{\theta_E} \left\{ \sum_{l=0}^{\infty} \sum_{s=-l}^l \left(H_{ls}^{(l,k)} \left(\frac{2l+1}{\hat{i}_l(\kappa a^{(l,k)})} - e^{-\kappa a^{(l,k)}} \hat{k}_l(\kappa a^{(l,k)}) \right) + F_{ls}^{(l,k)} + XH_{ls}^{(l,k)} \right) Y_{ls}^{(l,k)}(\theta', \phi') \right\} \overline{Y_{nm}^{(l,k)}}(\theta', \phi') \sin \theta' d\theta' d\phi' \\ &= \sum_{l=0}^{\infty} \sum_{s=-l}^l I_{E,lsnm}^{(l,k)} \left(H_{ls}^{(l,k)} \left(\frac{2l+1}{\hat{i}_l(\kappa a^{(l,k)})} - e^{-\kappa a^{(l,k)}} \hat{k}_l(\kappa a^{(l,k)}) \right) + F_{ls}^{(l,k)} + XH_{ls}^{(l,k)} \right) \end{aligned} \quad (20)$$

where I_E , the exposed surface integral matrix, is computed using quadrature method with M_{grid} uniform surface grid points

$$I_{E,lsnm}^{(l,k)} \equiv \frac{1}{4\pi} \int_{\phi_E} \int_{\theta_E} Y_{ls}^{(l,k)}(\theta', \phi') \overline{Y_{nm}^{(l,k)}}(\theta', \phi') \sin \theta' d\theta' d\phi' \approx \frac{1}{M_{\text{grid}}} \sum_{k=1}^{M_E} Y_{ls}^{(l,k)}(\theta_k, \phi_k) \overline{Y_{nm}^{(l,k)}}(\theta_k, \phi_k) \quad (21)$$

A similar transform to eq 20 can be written for $F_{nm}^{(l,k)}$. Finally, we truncate the series at pole order p to get the iterative equations

$$\frac{H_{nm}^{(l,k)}}{\hat{i}_n(\kappa a^{(l,k)})} = \sum_{l=0}^p \sum_{s=-l}^l I_{E,lsnm}^{(l,k)} \left(H_{ls}^{(l,k)} \left(\frac{2l+1}{\hat{i}_l(\kappa a^{(l,k)})} - e^{-\kappa a^{(l,k)}} \hat{k}_l(\kappa a^{(l,k)}) \right) + F_{ls}^{(l,k)} + XH_{ls}^{(l,k)} \right) \quad (22a)$$

$$F_{nm}^{(I,k)} = \sum_{l=0}^p \sum_{s=-l}^l I_{E,lsnm}^{(I,k)} (e^{-\kappa a^{(I,k)}} [\hat{k}_l(\kappa a^{(I,k)}) - (2l+1)\hat{k}_{l+1}(\kappa a^{(I,k)})] H_{ls}^{(I,k)} + (2l+1-l\epsilon) F_{ls}^{(I,k)} + X F_{ls}^{(I,k)}) \quad (22b)$$

Equations 22a and 22b, along with eqs 14a and 14b, represent a key result of this paper. The equations are iteratively evaluated, until the values of $\mathbf{F}^{(I,k)}$ and $\mathbf{H}^{(I,k)}$ converge to a stipulated tolerance. The operations are simply matrix-vector multiply, $\mathbf{y} = \mathbf{A}\mathbf{x}$, where the vector \mathbf{x} is constantly updated using the latest values of $\mathbf{F}^{(I,k)}$ and $\mathbf{H}^{(I,k)}$. During computation, the surface integral coefficients $I_{E,lsnm}^{(I,k)}$, and fixed charge coefficients $E_{nm}^{(I,k)}$ and $LE_{nm}^{(I,k)}$ are precomputed for each sphere (I,k) prior to simulation, while $LF_{nm}^{(I,k)}$, $LH_{nm}^{(I,k)}$, and $LHN_{nm}^{(I,k)}$ are updated via multipole-to-local operations (see implementation section below).

In summary, our approach to solve the LPBE is as follows: (1) For each sphere k in molecule I , we apply the addition theorems to express the potentials $\Phi_{\text{in}}(\mathbf{r})$ and $\Phi_{\text{out}}(\mathbf{r})$ as spherical harmonic expansions containing unknown coefficients ($F_{nm}^{(I,k)}$ and $H_{nm}^{(I,k)}$) representing sphere k 's polarization charges. (2) We impose boundary conditions on the sphere surface to derive boundary equations. (3) We account for charges from other spheres and molecules by re-expanding their polarization coefficients ($F_{nm}^{(J,l)}$ and $H_{nm}^{(J,l)}$) about the center of sphere k using “multipole-to-local” operations. (4) We then solve the boundary equations for $F_{nm}^{(I,k)}$ and $H_{nm}^{(I,k)}$ iteratively using a novel fast iterative method (*inner-iteration*). (5) We repeat steps 1–4 for all other spheres (*outer-iteration*) until the convergence criteria is reached.

Convergence is monitored using relative change in $\mathbf{H}^{(I,k)}$ between the t th and $(t-1)$ th outer iterations

$$\mu_{H,t}^{(I,k)} \equiv \frac{\sum_{n=0}^p \sum_{m=-n}^n |H_{nm,t}^{(I,k)} - H_{nm,t-1}^{(I,k)}|}{\frac{1}{2} \sum_{n=0}^p \sum_{m=-n}^n |H_{nm,t}^{(I,k)}| + |H_{nm,t-1}^{(I,k)}|} \quad (23)$$

We now can calculate the interaction energies from converged values of \mathbf{H} . The interaction energy of sphere k is the inner product of its effective charge distribution with the potential due to external sources. The interaction energy $W^{(I)}$ of each molecule I is the sum of interaction energies of its constituent spheres

$$W^{(I)} = \sum_{k=1}^{N_s^{(I)}} \langle \mathbf{LHN}^{(I,k)}, \mathbf{H}^{(I,k)} \rangle = \sum_{k=1}^{N_s^{(I)}} \sum_{n=0}^p \sum_{m=-n}^n LHN_{nm}^{(I,k)} \bar{H}_{nm}^{(I,k)} \quad (24)$$

Implementation of Re-expansion Operations. To solve for $\mathbf{F}^{(I,k)}$ and $\mathbf{H}^{(I,k)}$, we need to first account for the polarization charges from all other spheres via $\mathbf{LF}^{(I,k)}$, $\mathbf{LH}^{(I,k)}$, and $\mathbf{LHN}^{(I,k)}$. To do this, we convert source multipoles \mathbf{F} and \mathbf{H} from other spheres to target local expansions centered at $\mathbf{c}^{(I,k)}$. If the source and target spheres are well-separated (see criterion below), the re-expansion can be accomplished

analytically through multipole-to-local operators \mathbf{T}_0 and \mathbf{T}_κ . The procedure for computing coefficients of \mathbf{T}_0 and \mathbf{T}_κ has been previously detailed in ref 1. For *intra-molecular* re-expansions (i.e., from spheres j to center of sphere k in the same molecule I)

$$\mathbf{LF}^{(I,k)} = \sum_{j \neq k}^{N_s^{(I)}} \mathbf{T}_0^{(I,k)(I,j)} \mathbf{F}^{(I,j)}; \quad \mathbf{LH}^{(I,k)} = \sum_{j \neq k}^{N_s^{(I)}} \mathbf{T}_\kappa^{(I,k)(I,j)} \mathbf{H}^{(I,j)} \quad (25)$$

or *intermolecular* re-expansions (i.e., from spheres l to center of sphere k in the same molecule I)

$$\mathbf{LHN}^{(I,k)} = \sum_{J \neq I}^{N_{\text{mol}}} \sum_{l=1}^{N_s^{(J)}} \mathbf{T}_\kappa^{(I,k)(J,l)} \mathbf{H}^{(J,l)} \quad (26)$$

The analytical re-expansion operators are only valid when the target center $\mathbf{c}^{(I,k)}$ lies outside the bounding sphere of the source charge distribution, so they cannot be used in cases where source and target spheres overlap. Nonetheless, the local expansions $\mathbf{LF}^{(I,k)}$ and $\mathbf{LH}^{(I,k)}$ are still well-defined and could be directly computed using discrete versions of eqs 8c and 10b: a procedure we termed “numerical re-expansion”, as described below. To our knowledge this method of circumventing the restriction by analytical re-expansion has not been previously documented.

We first discretize the surface of source sphere j uniformly into M_p patches, with each patch b centered at $\mathbf{r}_b^{(I,j)} = [a^{(I,j)}, \theta_b^{(I,j)}, \phi_b^{(I,j)}]$. We then compute the surface charge on the b th patch $\tilde{q}_b^{(I,j)} = 4\pi \tilde{q}^{(I,j)}(\theta_b^{(I,j)}, \phi_b^{(I,j)})/M_p$, where $\tilde{q}^{(I,j)} = \tilde{f}^{(I,j)}$ or $\tilde{h}^{(I,j)}$ from eqs 15a and 15b. The local expansions of sphere j 's multipoles recentered on k are then approximated from eqs 8c and 10b as

$$LF_{nm}^{(I,k)} \approx \sum_{b=1}^{M_p} \frac{f_b^{(I,j)}}{r_b^{(I,k)}} \left(\frac{a^{(I,k)}}{r_b^{(I,k)}} \right)^n \bar{Y}_{nm}(\theta_b^{(I,k)}, \phi_b^{(I,k)}) \quad (27a)$$

$$LH_{nm}^{(I,k)} \approx \sum_{b=1}^{M_p} \frac{h_b^{(I,j)}}{r_b^{(I,k)}} \left(\frac{a^{(I,k)}}{r_b^{(I,k)}} \right)^n e^{-\kappa r_b^{(I,k)}} \hat{k}_n(\kappa r_b^{(I,k)}) \bar{Y}_{nm}(\theta_b^{(I,k)}, \phi_b^{(I,k)}) \quad (28b)$$

where $\mathbf{r}_b^{(I,k)} = \mathbf{r}_b^{(I,j)} - (\mathbf{c}^{(I,k)} - \mathbf{c}^{(I,j)})$. The re-expansion becomes exact as M_p approaches infinity, although in practice we find that a value of $M_p \approx 2.5p^2$ adequately captures features of the surface charge distributions. Numerical re-expansion is also used in cases where the source and target spheres are nonoverlapping but not well-separated, which we defined as when the distance between sphere surfaces is less than 5 Å. At such short distance, analytical re-expansion requires a high number of poles for a stipulated level of error. Since both computational time and memory for \mathbf{T} scales with p^3 it is more efficient to perform the re-expansion using direct numerical method.

We have also derived a formula using Greengard's error bound³³ to adaptively determine the minimum pole order adequate for a re-expansion operation. To re-expand sphere (J,j) 's multipole to a local expansion at target center (I,k) within an error of ϵ_x , the pole order required is given by

$$p = \log \left(\frac{\sum_{\text{charges on } j} |\tilde{q}_j|}{\epsilon_X a^{(j)} (c - 1)} \right) / \log(c) - 1 \quad (29)$$

where $c = (|\mathbf{c}^{(I,k)} - \mathbf{c}^{(J,j)}|) / (a^{(I,j)} - 1)$ and $\tilde{q} = \tilde{f}$ or \tilde{h} are the surface polarization charges. The optimal pole order is calculated on the fly every outer iteration.

Further Implementation Details. The surface integral coefficients $I_{E,lsnm}^{(I,k)}$ involve numerical quadratures that are precomputed for each sphere (I,k) ; we have found that the number of quadrature points should scale with pole number as $M_{\text{grid}} \approx 20p^2$, which we found to be adequate for capturing the spatial features of the integrand in eq 21.

To prepare a target molecule for computation, we must discretize it into a collection of overlapping spheres. To do so, we first convert its PDB file to PQR format using the PDB2PQR webserver.^{12,13} We then obtain its solvent excluded surface (SES) using MSMS³⁴ and a chosen probe radius r_p in Å. We proceed with a Monte Carlo search algorithm to find the minimum number of spheres and corresponding radii that satisfying the following criteria: (i) The sphere surface must be at least d (in Å) away from the outermost atom center. The distance d can be held constant or set to the van de Waals radius of each atom. (ii) The surface of the spheres cannot protrude more than t (in Å) from the SES surface. The search is terminated when each atom is encompassed by at least one sphere.

Finally, the code is implemented in C++ and is parallelized in a shared memory framework using openMP 2.0. Timings for PB-SAM and APBS for test cases in Results are based on single processor runs on an Intel(R) Xeon(R) CPU 2.27 GHz node with 24GB of physical memory; we did this to compare PB-SAM in a serial version against the APBS serial code. Parameters used for all APBS calculations are available in Supporting Information. Timings for Brome Mosaic virus calculations are performed on the same node using 8 processors.

Results

Nonoverlapping Spherical Test Cases. We first assess the accuracy of PB-SAM and APBS finite difference solutions against analytical values for three test systems involving 2, 27, and 343 nonoverlapping spherical dielectric cavities (of diameter 20, 15, and 5 Å, respectively) with internal charges placed near the dielectric boundaries (Table 1). For large spheres, this corresponds to a highly asymmetric charge arrangement, while as sphere size decreases the charge

Table 2. Comparison of APBS against the Analytical Model for Test Systems Described in Table 1

test system	grid points	resolution (Å)	run time (s)	memory (GB)	overall relative error	maximum relative error
1	65 × 65 × 65	1.5625	3	0.08	19.7%	34.8%
1	129 × 129 × 129	0.7813	29	0.47	14.4%	24.7%
1	257 × 257 × 257	0.3906	142	3.50	11.2%	31.7%
1	513 × 513 × 513	0.1953	1315	27.8	4.9%	11.4%
2	513 × 513 × 513	0.1953	1216	27.8	1.9%	5.3%
3	513 × 513 × 513	0.1953	1421	27.8	1.1%	4.9%

distribution approaches a monopole. The exact analytical solution of the PBE for multiple nonoverlapping spheres has only become available recently.¹ In all cases, the salt concentration is set to 0.05 M, corresponding to $\kappa = 0.07374$. We chose a low salt concentration to show a worse case scenario for computational timings that cannot exploit an aggressive interaction cutoff that would be legitimate at higher physiological salt concentrations. Convergence is reached when the relative change $\mu_{H,i}^{(I,k)}$ falls below 10^{-2} for all spheres.

For test system 1 (two nonoverlapping spheres), we computed the APBS solutions at four different grid resolutions (0.19–1.56 Å) that are typically used in biomolecular applications, and compared the potential value over the entire surface against the analytical model, as well as reporting the corresponding memory requirements and timings (Table 2). The APBS timing scales linearly with the number of grid points because does the memory cost that largely reached the limit of 27 GB on our computing node at the highest resolution we tested. At the most coarse resolution, we find that the APBS error can be as high as $\sim 20\%$ of the theoretical result: as the APBS grid spacing decreases the APBS accuracy increases, reaching $\sim 5\%$ of the true value. The range of our computed errors for APBS agree with the error analysis performed by Moy et al.,³⁵ who found that the errors from using finite difference methods could range from a few percent at 0.5 Å to more than 100% at 2 Å. Given that these commonly used grid resolutions entailed such large errors, we feel that more systematic benchmarking should be done in the future to quantify accuracies in numerical Poisson–Boltzmann solutions to ascertain the impact on force and free energy computations. Using the highest resolution grid but increasing the number of spheres in test systems 2 and 3, the APBS solution gets corresponding better as the charge distribution simplifies, with average errors of $\sim 2\%$ and $\sim 1\%$, respectively.

Table 3 shows that the corresponding results for our PB-SAM model. For each system, we report the PB-SAM results

Table 1. Spherical Test Systems for Comparison of APBS and PB-SAM to Analytical Model Solution in Tables 2 and 3^a

test system	description	charge configuration [position from center], charge [e]
1	2 dielectric cavities of radius 20 Å	cavity 1 [18, 0, 0], +3 cavity 2 [−18, 0, 0], −3
2	27 dielectric cavities of radius 15 Å	all cavities [13, 0, 0], +1; [−13, 0, 0], −1 [0, 13, 0], +2; [0, −13, 0], −2 [0, 0, 13], +1; [0, 0, −13], −1
3	343 dielectric cavities of radius 5 Å	all cavities [3, 0, 0], +1; [−3, 0, 0], −1 [0, 3, 0], +2; [0, −3, 0], −2 [0, 0, 3], +1; [0, 0, −3], −1

^a Cavities have surface-to-surface separation of 1 Å from one another.

Table 3. Comparison of PB-SAM against Analytical Model for Test Systems Described in Table 1

test system	number of multipoles	run time (s)	memory (GB)	overall relative error	maximum relative error
1	30	4.3	0.023	13.5%	17.6%
1	35	12.1	0.031	4.3%	4.6%
1	40	20.7	0.051	2.4%	1.9%
2	10	1.4	0.015	13.6%	26.7%
2	15	2.3	0.021	6.4%	11.8%
2	20	7.6	0.033	2.2%	4.1%
2	30	46.5	0.082	0.4%	4.4%
3	5	22.2	0.108	4.4%	9.6%
3	10	28.4	0.167	0.1%	0.3%

for various pole orders, to demonstrate how a user would be able to tune into a desired level of accuracy in PB-SAM's using the pole order. From Table 3, it is clear that we can quickly exceed the accuracy of the APBS solution at a fraction of the cost and memory requirements for all three systems. In all three test cases, very few poles ($p \leq 40$) are needed to define a high accuracy solution, primarily because there are no problematic deep cusp dielectric geometries in the nonoverlapping sphere case.

Overlapping Spherical Test Cases. Our second comparison involves two overlapping spheres of various sizes. In this case, no analytical solution is known, but we can define a benchmark calculation based on a high quality PB-SAM solution computed at $p = 140$ and $M_p = 200\,000$ (PB-SAM140). The one-time precomputation of the surface integral matrix (I_E) scales with p^4 both in timing and storage. We stopped generating I_E at 140 poles, which took 48 h per sphere. Before using $p = 140$ as a benchmark, we have confirmed that the overall relative difference between potential solved using $p = 140$ and lesser pole orders decays with increasing pole order: the overall relative difference between the $p = 130$ and $p = 140$ solutions are less than 0.5%.

In Table 4, we compare the relative difference in surface potential against PB-SAM140 as sphere size increases. We considered the worst-case scenario by placing the positive charge close to the surface, at a fixed distance of 1.73 Å

below the cusp region, so that as sphere size grows it results in higher asymmetry of the charge distribution. For each sphere radius, we first compared the surface potential computed by APBS against that of PB-SAM140 and then perform the same comparison between PB-SAM of various pole orders against PB-SAM140. For APBS at a maximum grid dimension fixed at 513,³ as the system size increases the grid resolution and accordingly the solution accuracy deteriorate. On the other hand, the accuracy of PB-SAM is a function of the pole order and the size of constituent spheres (sphere resolution). Table 4 thus provides a handy estimate of PB-SAM error to help guide our choice of sphere resolutions and pole order for subsequent application to realistic protein cases. For the two overlapping sphere case, PB-SAM with 60 poles is able to achieve relative errors comparable to APBS with comparable total solve time, and with less memory requirements. We want to point out that the total solve time for PB-SAM reported in Table 4 is principally dominated by the one-time cost of surface integral computation (1140 s), while the actual time for solving the iterative equations, eqs 22a and 22b, are between 9 s and 2 min.

It is interesting to note that, for a fixed number of poles, PB-SAM's relative error increases with increasing sphere sizes. Since the boundary equations are formulated and solved in scaled representations, they should be independent of sphere sizes. The two potential sources of error are if M_p is insufficient in discriminating the positions of the source surface charges for numerical re-expansion, or the scaled fixed charge multipole $E_{nm}^{(l,k)}$ decays more slowly with poles with increasing charge asymmetry. While we found that increasing M_p by a factor of 40 resulted in no change in potential, the term $(r_a/a^{(l,k)})^n$ in $E_{nm}^{(l,k)}$ converges slower at large sphere radii; hence, more poles are needed to describe the corresponding increase in charge asymmetry. In practice, charges in realistic biomolecules are more evenly distributed; hence, their fixed charge multipoles will converge much faster. The convergence improves further when smaller spheres are used to define higher resolution dielectric

Table 4. Two Overlapping Spheres with Varying Sphere Sizes^a

sphere size	APBS					PB-SAM				
	grid size (Å)	solve time (s)	memory (GB)	relative error	maximum relative error	pole order	solve time (s)	memory (GB)	relative error	maximum relative error
2	0.0195	960	27.8	0.6%	1.6%	20	1141	0.018	12.1%	16.1%
						30	1143	0.030	7.0%	9.2%
						40	1149	0.057	3.8%	5.6%
						60	1209	0.230	1.5%	2.1%
5	0.0391	1018	27.8	1.6%	3.5%	20	1141	0.018	14.4%	20.6%
						30	1143	0.030	7.7%	10.5%
						40	1148	0.057	5.5%	7.8%
						60	1315	0.229	2.3%	3.9%
15	0.1172	1,158	27.8	4.6%	9.7%	20	1141	0.018	21.7%	30.4%
						30	1143	0.030	12.5%	18.4%
						40	1148	0.057	8.6%	14.4%
						60	1223	0.229	4.3%	6.9%
50	0.3906	1,276	27.8	16.8%	32.8%	20	1141	0.018	41.8%	36.2%
						30	1142	0.030	27.7%	25.3%
						40	1148	0.057	19.6%	18.2%
						60	1180	0.229	11.3%	13.3%

^a Comparison of the surface potential computed with APBS and PB-SAM ($M_{\text{grid}} = 100\text{k}$, $M_p = 2.5\rho^2$) against PB-SAM140.

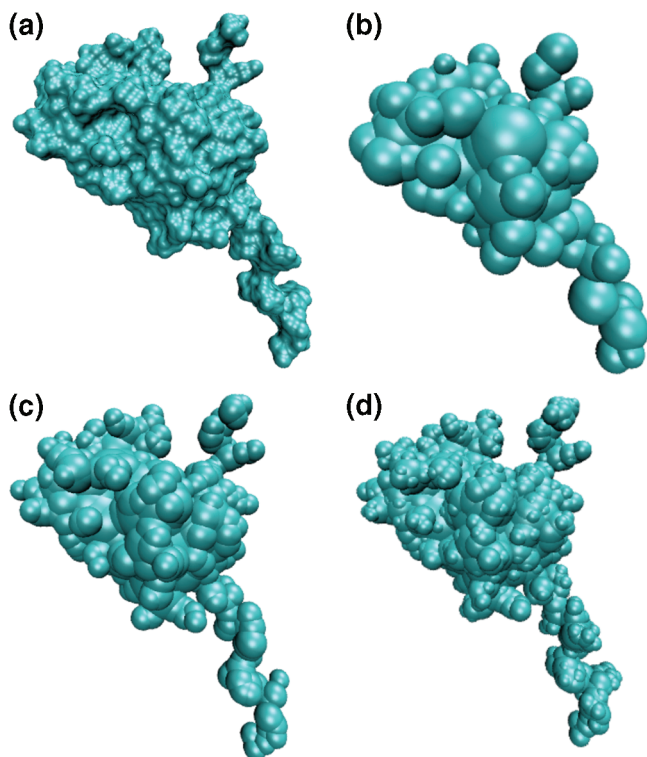


Figure 3. Representations of 1YC6 monomer based on different discretization criteria: (a) the solvent-excluded surface computed using MSMS with $p = 1.4 \text{ \AA}$, (b) 107 spheres with $p = 1 \text{ \AA}$, $d = 1 \text{ \AA}$, $t = 2 \text{ \AA}$, (c) 354 spheres with $p = 1 \text{ \AA}$, $d =$ atomic vdW radii, $t = 1 \text{ \AA}$, and (d) 712 spheres with $p = 1 \text{ \AA}$, $d =$ atomic vdW radii, $t = 0.5 \text{ \AA}$.

boundaries. Hence Table 4 shows that PB-SAM's relative error decreases with smaller spheres and higher pole, and our simplified case with maximum charge asymmetry provides a worse case upper bound on the relative error for $20 < p < 60$. This will inform estimates of error in our calculation of the bromine mosaic virus in the following section.

Bromine Mosaic Virus. We have also applied our PB-SAM method to solve for the potential around a biological molecule, the $T = 1$ particle of the bromine mosaic virus (BMV) capsid (PDB code 1YC6). The virus has been shown to convert from $T = 3$ (comprising of 180 monomers) to $T = 1$ (comprising of 60 monomers) capsid under proteolytic conditions.³⁶ Each capsid protein monomer is comprised of 154 amino acids. To prepare the PDB file for calculation,

we converted chain A of the PDB file into PQR format using the PDB2PQR server,^{37,38} which also assigned partial atomic charges using the AMBER 99 force field.³⁹ We then discretized the protein into a collection of overlapping spheres using an in-house algorithm (see implementation details in Methods). Using discretization criteria that varies in spatial resolution, we generated three representations of the protein monomer with 107, 354, and 712 spheres, and Figure 3 compares the dielectric boundary representation against the solvent excluded surface computed using MSMS^{34,40} with probe radius $r_p = 1.4 \text{ \AA}$. The generation of 107, 354, and 712 spheres took 11, 30, and 43 min, respectively, which is typically a one-time cost if the dielectric representation does not change during the course of a Brownian dynamics simulation, for example. The resulting rendered dielectric boundary in each case were then used to generate a corresponding APBS solution on the maximum allowed grid points of 513^3 given our maximum memory of 24GB.

Table 5 describes the computational time and memory resources for PB-SAM to calculate the self-polarization of one 1YC6 monomer, and the mutual polarization of an array of 60 monomers that make up the unassembled BMV capsid. Since it is our intention to study the dynamics of BMV capsid assembly via Brownian dynamics in future work, we consider the breakdown of computational cost and memory as (1) a one time cost to prepare the surface integral of the chosen dielectric representation of the 1YC6 monomer, (2) the one-time cost to self-polarize each monomer, and (3) the cost to mutually polarize the 60 monomers. In the context of a Brownian dynamic simulation, Table 5 represents the cost of the initialization phase that will require "cold" guesses for \mathbf{F} and \mathbf{H} for steps 2 and 3, and the timings will be nonoptimal relative to later solutions that will provide better initial guesses as the dynamics algorithm proceeds as the capsid assembles.

The PB-SAM computational cost depends on the number of poles and number of spheres, and timings are faster or slower depending on how much of the calculation can be done in memory. We use Table 4 to guide our choices of pole order and sphere resolution. We will focus our PB-SAM solutions at a $\sim 5\text{--}7\%$ error by choosing pole order $20 < p < 60$ and keeping average size of spheres of the dielectric boundary representation between $2\text{--}5 \text{ \AA}$. For step 1, Table 5 shows the one time surface integral

Table 5. Computational Timing and Memory Resources Using PB-SAM for Capsid Assembly^a

number and median sphere radius	poles	time to calculate surface integrals (s)	self-polarization		mutual-polarization	
			time (s)	memory (GB)	time (s)	memory (GB)
107 spheres 4.40 Å	40	1083	280	3.6	2589	4.4 ^b
	50	4131	552	7.2		
	60	12336	1180	13.3		
354 spheres 3.06 Å	30	423	603	7.8	9365	13.5 ^b
	40	2380	2091	7.1 ^b		
	50	9079	4934	17.1 ^b		
712 spheres 1.91 Å	20	70	271	8.6	16046	33.8 ^b
	30	802	1177	17.5		
	40	4508	3707	14.2 ^b		

^a Self-polarization of 1YC6 monomer and mutual polarization of 60 monomers of BMV capsid for various dielectric representations (Figure 3). ^b Memory-saving mode.

cost of the 1YC6 monomer, which scales as $O(M_{\text{grid}}p^4)$ (see Methods), varies between several minutes to several hours. However, a nice benefit is that as resolution increases the sphere size, and hence, M_{grid} decreases as do the number of needed poles, which together mitigates the time of calculating more spheres. The cost to self-polarize will depend on the available memory; in memory-saving mode the re-expansion operators T_0 and $T_{\mathbf{k}}$ are computed on the fly, instead of being stored in memory and hence increase the cost of the calculation. In Table 5, the self-polarization timings are based on a “cold” guess of $F^{(j,k)} = 0$ and $H^{(j,k)}$ approximated using the fixed charges, and iterated until the relative change in $H^{(j,k)}$ falls below 10^{-2} for all spheres.

Unlike the idealized test cases in Table 1, we do not have a benchmark “exact” solution for the 1YC6 monomer, since no analytical solution exists for nonspherical geometries, and computation of hundreds of surface integral matrices to $p = 140$ for PB-SAM is currently intractable. Instead we can obtain error estimates of APBS and PB-SAM by looking up corresponding PBE solution parameters of grid size for APBS and pole and sphere size for PB-SAM in Table 4. For the 1YC6 monomer, the APBS result is necessarily evaluated at a low resolution of 0.22 \AA on the basis of the maximum allowed grid points of 513^3 given our maximum memory of 24 GB; Table 4 suggests that the APBS relative error would be $\sim 10\text{--}12\%$ for this system. For PB-SAM, the error estimate from Table 4 for spheres between $2\text{--}5 \text{ \AA}$ solved to pole order $20 < p < 60$, fall between 2.1 to 20.6% . Any direct quantitative comparison between the errors of the APBS and PB-SAM solutions is not possible since there is no exact benchmark for this case, and we can only provide the numerical *difference* between methods (which is $\sim 10\%$) and not *error* between the two numerical solutions. However, using our error estimate from Table 4, it suggests that with $30\text{--}40$ poles for the representations of 107 and 353 spheres and $20\text{--}30$ poles for 712 spheres, PB-SAM is a higher quality solution at a comparable CPU cost and memory of the APBS solution.

Finally, we have evaluated the potential of an assembly of 60 copies of 1YC6 monomers in a $5 \times 4 \times 3$ array, corresponding to a system size of $165 \text{ \AA} \times 220 \text{ \AA} \times 275 \text{ \AA}$ (Figure 4). The array configuration is intended to mimic late stage assembly, at which the entire capsid system is compact and mutual polarization becomes significant and more difficult to converge (as opposed to the 60 monomers being well separated). All monomers were given the same initial guess of F^{self} and H^{self} from the converged self-polarization step, and the computational time and memory to calculate the total (self- and mutual) polarization is given in Table 5. The memory for the 712-sphere representation required 33 GB of virtual memory, which is not as efficient if it were able to fit in the available 24 GB of physical memory. The fact that the calculation of a high quality solution is doable on a single standard commodity node

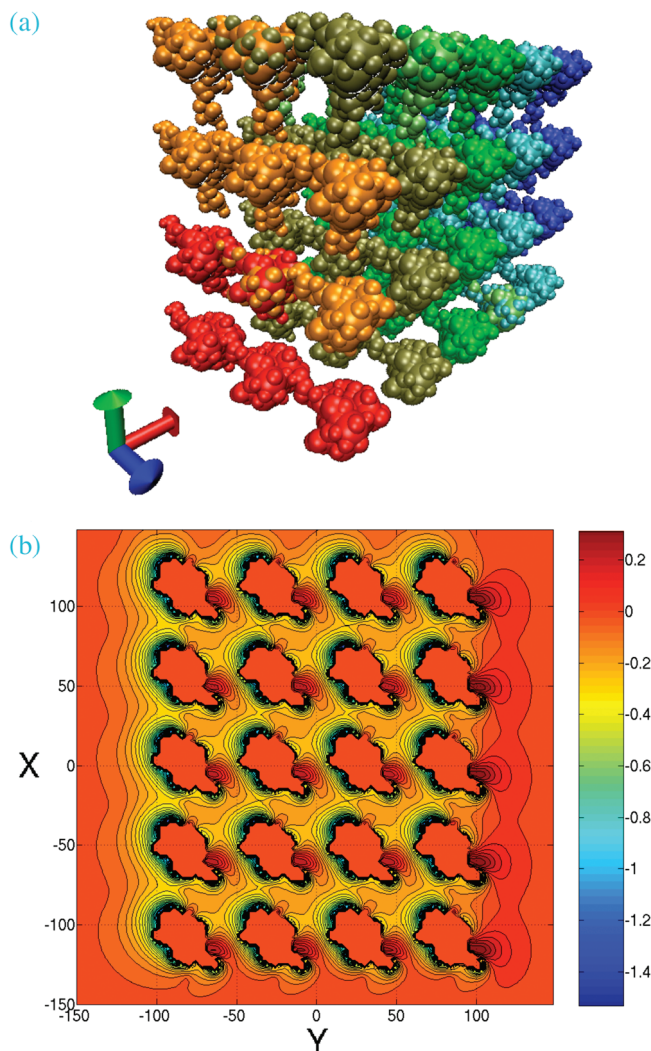


Figure 4. Array of 60 virus monomers: (a) array configuration and (b) potential profile of a cross-section through the $z = 0$ plane with twenty monomers. Contour lines at 0.05 kT .

is a strength of the PB-SAM approach, although further optimization will be explored in the future.

Conclusion

We have developed a novel method for solving the linearized Poisson–Boltzmann equation by discretizing the protein surface as a collection of spheres, in which the surface charges can be iteratively solved by our recent analytical solution of the PBE equations for spherical geometries in which mutual polarization is treated exactly.¹ We have compared PB-SAM and the finite difference PB solver APBS against two new benchmarks never before available to compare numerical methods. First, we show that PB-SAM converges to the analytical solution of hundreds of spheres with better accuracy and at greatly reduced cost relative to APBS. Second, the PB-SAM solution using 140 poles allows us to define a high quality benchmark to describe the electrostatic potential for two overlapping spheres that are models for cusp-like features of protein active sites, in which we show that our PB-SAM solution converges to the correct solution with the same computational cost or better than the finite difference solution. Finally we illustrate the strength

of the PB-SAM approach by computing the potential profile of a close configuration of 60 T1-particle forming monomers of the bromine mosaic virus (PDB code 1YC6), with clear improvements in accuracy relative to other numerical PB solutions, given a fixed hardware configuration of physical memory.

Further development is necessary to enable PB-SAM's application in large-scale Brownian dynamic simulations. The current version of PB-SAM expends significant computational time solving eqs 22a and 22b iteratively. This step was implemented simply as repeated calls to the BLAS matrix-vector multiply routine *dgemv* but can be accelerated by preconditioning eqs 22a and 22b and using a more sophisticated linear system solving method, such as generalized minimal residual method. We also noted during our benchmarking studies that when our current convergence criterion is relaxed, the resulting surface potential is unchanged, so there is room to explore a less stringent but adequate convergence criterion. Finally, forces and torques are required for Brownian dynamic simulation. We have derived in reference¹ how forces and torques can be computed analytically for spherical dielectrics. The same formulation can be extended to the overlapping sphere representation in PB-SAM via superposition, which is ongoing work in our lab.

Acknowledgment. We gratefully acknowledge support from an NIH Multiscale grant and NERSC for computational resources, and Dr. Itay Lotan for providing source code for ref 1 and clarifications.

Supporting Information Available: Parameters used for all APBS calculations. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- Lotan, I.; Head-Gordon, T. *J. Chem. Theory Comput.* **2006**, *2*, 541.
- Davis, M. E.; Mccammon, J. A. *Chem. Rev.* **1990**, *90*, 509.
- Chapman, D. L. *Philos. Mag., Ser. 6* **1913**, *25*, 475.
- Gouy, G. *J. Physique* **1910**, *9*, 456.
- Debye, P.; Huckel, E. *Phys. Z.* **1923**, *24*, 185.
- Derjaguin, B.; Landau, L. *Prog. Surf. Sci.* **1993**, *43*, 30.
- Verwey, E. J. W. *Philips Res. Rep.* **1945**, *1*, 33.
- Kirkwood, J. G. *J. Chem. Phys.* **1934**, *2*, 351.
- Fenley, A. T.; Gordon, J. C.; Onufriev, A. *J. Chem. Phys.* **2008**, *129*.
- McClurg, R. B.; Zukoski, C. F. *J. Colloid Interface Sci.* **1998**, *208*, 529.
- Phillies, G. D. *J. Chem. Phys.* **1974**, *60*, 983.
- Sader, J. E.; Lenhoff, A. M. *J. Colloid Interface Sci.* **1998**, *201*, 233.
- Lu, B. Z.; Zhou, Y. C.; Holst, M. J.; McCammon, J. A. *Commun. Comput. Phys.* **2008**, *3*, 973.
- Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037.
- Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435.
- Rocchia, W.; Alexov, E.; Honig, B. *J. Phys. Chem. B.* **2001**, *105*, 6507.
- Zhou, Z. X.; Payne, P.; Vasquez, M.; Kuhn, N.; Levitt, M. *J. Comput. Chem.* **1996**, *17*, 1344.
- Davis, M. E.; Madura, J. D.; Luty, B. A.; Mccammon, J. A. *Comput. Phys. Commun.* **1991**, *62*, 187.
- Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; Mccammon, J. A. *Comput. Phys. Commun.* **1995**, *91*, 57.
- Chen, L.; Holst, M. J.; Xu, J. C. *SIAM J. Numer. Anal.* **2007**, *45*, 2298.
- Holst, M.; Baker, N.; Wang, F. *J. Comput. Chem.* **2001**, *22*, 475.
- Zhou, H. X. *Biophys. J.* **1993**, *65*, 955.
- Bordner, A. J.; Huber, G. A. *J. Comput. Chem.* **2003**, *24*, 353.
- Boschitsch, A. H.; Fenley, M. O.; Zhou, H. X. *J. Phys. Chem. B* **2002**, *106*, 2741.
- Juffer, A. H.; Botta, E. F. F.; Vankeulen, B. A. M.; Vanderploeg, A.; Berendsen, H. J. C. *J. Comput. Phys.* **1991**, *97*, 144.
- Lu, B. Z.; Cheng, X. L.; Huang, J. F.; McCammon, J. A. *J. Chem. Theory Comput.* **2009**, *5*, 1692.
- Lu, B. Z.; Cheng, X. L.; Huang, J. F.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 19314.
- Lu, B. Z.; McCammon, J. A. *J. Chem. Theory Comput.* **2007**, *3*, 1134.
- Zauhar, R. J.; Morgan, R. S. *J. Mol. Biol.* **1985**, *186*, 815.
- Gumerov, N. A.; Duraiswami, R. *SIAM J. Sci. Comput.* **2003**, *25*, 1344.
- Abramowitz, M.; Stegun, I. A. *Handbook of Mathematical Functions*; U.S. Department of Commerce: Washington, D.C., 1968; Vol. 5.
- Chen, G.; Zhou, J. *Boundary Element Methods*, 1st ed.; Academic Press, London, San Diego, 1992.
- Cheng, H.; Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1999**, *155*, 468.
- Sanner, M. F.; Olson, A. J.; Spohner, J. C. *Biopolymers* **1996**, *38*, 305.
- Moy, G.; Corry, B.; Kuyucak, S.; Chung, S. H. *Biophys. J.* **2000**, *78*, 2349.
- Lucas, R. W.; Kuznetsov, Y. G.; Larson, S. B.; McPherson, A. *Virology* **2001**, *286*, 290.
- Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. *Nucleic Acids Res.* **2007**, *35*, W522.
- Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. *Nucleic Acids Res.* **2004**, *32*, W665.
- Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.
- Sanner, M. F. *J. Mol. Graphics Modell.* **1999**, *17*, 57.

Halogen-Ionic Bridges: Do They Exist in the Biomolecular World?

Peng Zhou,[†] Yanrong Ren,[§] Feifei Tian,^{‡,¶} Jianwei Zou,[¶] and Zhicai Shang^{*,†}

Department of Chemistry, Zhejiang University, Hangzhou 310027, China, Department of Biological and Chemical Engineering, Chongqing Education College, Chongqing 400067, China, College of Bioengineering, Chongqing University, Chongqing 400044, China, Key Laboratory for Molecular Design and Nutrition Engineering, Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, China, and Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611

Received March 27, 2010

Abstract: If considering that the pronouncedly charged halide anions are ubiquitous in the biological world, then it is interesting to ask whether the halogen-ionic bridges—this term is named by us to describe the interaction motif of a nonbonded halogen ion with two or more electrophiles simultaneously—commonly exist in biomolecules and how they contribute to the stability and specificity of biomolecular folding and binding? To address these problems, we herein present a particularly systematic investigation on the geometrical profile and the energy landscape of halogen ions interacting with and bridging between polar and charged molecular moieties in small model systems and real crystal structures, by means of ab initio calculation, database survey, continuum electrostatic analysis, and hybrid quantum mechanics/molecular mechanics examination. All of these unequivocally demonstrate that this putative halide motif is broadly distributed in biomolecular systems (>6000) and can confer a substantial stabilization for the architecture of proteins and their complexes with nucleic acids and small ligands. This stabilization energy is estimated to be generally more than 100 kcal·mol⁻¹ for gas-phase states or about 20 kcal·mol⁻¹ for solution conditions, which is much greater than that found in sophisticated water-mediated (<10 kcal·mol⁻¹) and salt (~ 3.66 kcal·mol⁻¹) bridges. In this respect, we would expect that the proposed halogen-ionic bridge, which has long been unrecognized in the arena of biological repertoires, could be appreciated in chemistry and biology communities and might be exploited as a new and versatile tool for rational drug design and bioengineering.

1. Introduction

Specific ion effects play an essential role in many physico-chemical and biological processes. Such effects exhibit a reoccurring trend called the Hofmeister series.¹ Originally,

it was thought that an ion's influence on macromolecular properties was caused at least in part by 'making' or 'breaking' bulk water structures.² Recent time-resolved and thermodynamic studies of water molecules in salt solutions, however, shed light on that, instead of remodeling water structures through ions, direct macromolecule–ion interactions as well as the interactions with water molecules that are bound to the macromolecules seem to be more responsible for the Hofmeister effect.³

In fact, the metal cation–protein/nucleic acid interactions, which are commonly known as coordinate bonding, have been well characterized in chemistry and biology communities. In contrast, the ubiquitous anions in biologi-

* Corresponding author. Telephone: +86-0571-87952379. Fax: +86-0571-87951895. E-mail: shangzc@zju.edu.cn.

[†] Department of Chemistry, Zhejiang University.

[§] Department of Biological & Chemical Engineering, Chongqing Education College.

[‡] College of Bioengineering, Chongqing University.

[¶] Department of Chemistry, University of Florida.

[¶] Ningbo Institute of Technology, Zhejiang University.

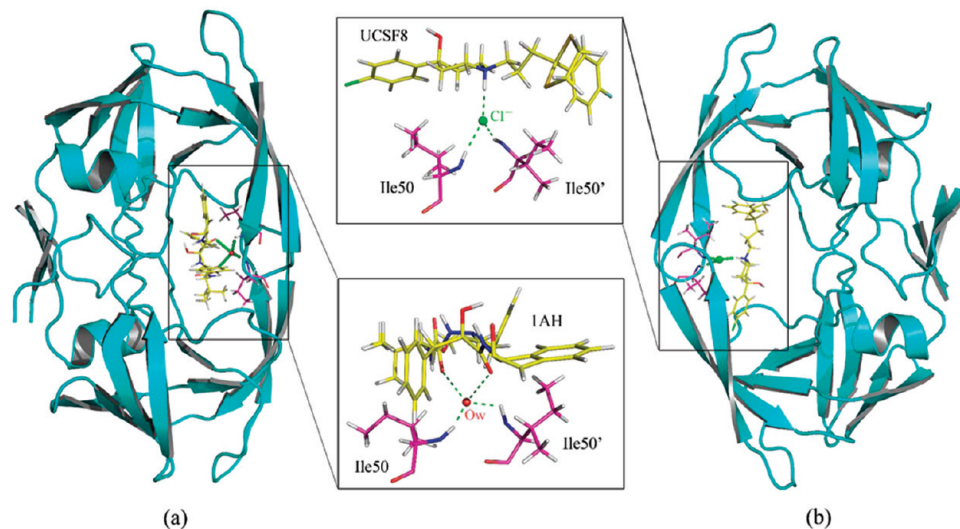


Figure 1. Crystallographic evidence showing that the role of a water molecule in mediating the hydrogen-bond network in biomolecules can be functionally replaced by a halogen ion. Usually, a conserved water molecule (Ow) is located in the active pocket of HIV-1 protease to mediate the hydrogen-bond network with its cognate substrates and noncognate inhibitors (PDB: 2cej) (a). However, there has an exception that a chlorine ion (Cl^-) is observed at the water's position in the complex of HIV-1 protease with its nonpeptide inhibitor UCSF8 (PDB: 1aid) (b).

cal systems, such as X^- (where $\text{X}^- = \text{F}^-, \text{Cl}^-, \text{Br}^-, \text{and } \text{I}^-$), SO_4^{2-} , and H_2PO_4^- , are traditionally recognized as counterions, and their interactions with biomolecules as well as the effect of these interactions on biomolecular functions have long been underappreciated in the field of biology. Theoretically, anions, especially the small, hard halogen ions, are expected to serve as a good hydrogen-bond acceptor and a friendly ion-pairing partner to specifically and nonspecifically interact with the polar hydrogen atoms and the basic groups of biomolecules. This point has been preliminarily rationalized by experimental and theoretical studies of simple molecular complexes in gas phase or vacuum conditions (see reviews).^{4,5} For example, the dissociation enthalpies of $\text{X}^- \cdots \text{H}_2\text{O}$ and $\text{CH}_3(\text{CH}_2)_n\text{OH} \cdots \text{X}^-$ adducts were early determined to be $\sim 10\text{--}30 \text{ kcal}\cdot\text{mol}^{-1}$ by using high-pressure mass spectrometry,^{6–9} which are even six-fold greater than the stabilization energy of a water dimer ($\sim 5 \text{ kcal}\cdot\text{mol}^{-1}$).¹⁰ The experimentally measured energies were later systematized by means of ab initio calculations using model systems.^{11–15}

Close contacts between the nonbonded halogen ions and the hydrogen atoms were observed in crystal structures of amino acids, peptides, and related molecules as early as 30 years ago.¹⁶ In the past two decades, the interactions of halogen ions with macromolecular systems, including colloids, polymers, and proteins, were investigated intensively via nuclear magnetic resonance (NMR),¹⁷ aqueous gel sieving chromatography,¹⁸ and vibrational sum frequency spectroscopy¹⁹ as well as molecular dynamics simulation.²⁰ Particularly, it was found that specific ion effects on protein stability could be explained by incorporating the ionic dispersion potentials into classical double-layer theory²¹ and that small anions, such as F^- , are prone to pair with charged groups, while larger anions, such as I^- , are more likely to be bound on hydrophobic patches of protein surfaces.²² Very recently and also very intriguingly, Heyda et al. have presented computational evidence for the ion-specific inter-

actions between biological entities and halides. By employing both nonpolarizable and polarizable force fields to simulate the dynamic behavior of amino acid–ion systems, they attained several clear trajectory pictures showing obvious congregations of halogen anions around the positively charged hydrogen atoms of basic amino acids.²³

It is known that water molecule can serve as mediator to “glue” adjacent polar groups together through hydrogen bonds and hydrophilic forces. Traditionally, these water-participating interactions are referred to as a water-mediated hydrogen-bond bridge²⁴ and a water-induced hydrophilic interaction.²⁵ Given that the halide anions, as mentioned above, are shown to be effective in interaction with biomolecules, a question would be raised naturally, that is, whether the halogen ions can bridge between the spatially vicinal moieties in biomolecules, just like what the water molecules do? In other words, do the putative halogen ion-participating interaction motifs, that we named halogen-ionic bridges to stress their shared similarities with water-mediated bridges, exist in the biomolecular world? Actually, there has been at least one crystallographic report clearly showing a Cl^- bound functionally between the residues Ile50/Ile50' of HIV-1 protease and the protonated tertiary amine of its nonpeptide inhibitor UCSF8, a haloperidol derivative which strongly inhibits both wild-type and mutant HIV-1 proteases (Figure 1b).²⁶ As we know, however, this Cl^- position is usually occupied by a conserved water molecule (Figure 1a).²⁷

To address these open questions related to the existence and significance of halogen-ionic bridges in biological context, in the present work we launch a systematic investigation on this putative halide motif through various theoretical and computational approaches. First, high-level quantum mechanical (QM) calculations were carried out for a series of small model systems to elucidate the geometrical preference and the energy landscape of halogen ions interacting with model molecules which mimic polar and charged

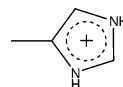
biomolecular groups. The resulting geometrical and energetic features of such interactions subsequently recurred in an exhaustive survey of the high-resolution crystal structures of all the biomolecules deposited in the Protein Data Bank (PDB),²⁸ including proteins, nucleic acids, and their complexes with small ligands. In particular, the electrostatic property and structural basis of halogen-ionic bridges in real biomacromolecular systems and their contributions toward the stability and specificity of protein architecture and protein–ligand recognition were analyzed in detail with the Poisson–Boltzmann model and a two-layer quantum mechanics/molecular mechanics (QM/MM) scheme. This study would provide solid evidence for the halogen-ionic bridges existing in and functionalizing to biomolecules and might give a new view to support the notion that direct ion–macromolecule interactions, rather than indirect water structures making and breaking by ions, are more responsible for the specific ion effects on biological systems.

2. Methods and Materials

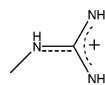
2.1. Quantum-Mechanical (QM) Calculation. The simplest model systems, i.e., water molecule (H₂O) and ammonium ion (NH₄⁺) in complex with four kinds of halogen ions (F[−], Cl[−], Br[−], and I[−]) and H₂O (serving as the neutral counterpart of halogen ions), were used to perform a detailed examination of geometrical, energetic, and electronic properties associated with the interactions of halogen ions with polar and charged groups, respectively. Potential energy surface scans were carried out at the Møller–Plesset second-order perturbation theory level,²⁹ in conjunction with the Dunning's augmented correlation consistent basis set, MP2/aug-cc-pVDZ. The equilibrium structures as well as corresponding parameters of atoms in molecules (AIM)³⁰ and natural bond orbitals (NBO)³¹ for these complexes were further obtained at the MP2/aug-cc-pVTZ level of theory; the more accurate intermolecular potentials for the equilibrium structures were evaluated using the coupled cluster with single, double, and noniterative triple excitations correction term, CCSD(T)/aug-cc-pVTZ. The supermolecule approach was employed to obtain intermolecular potentials (viz. $\Delta E_{\text{int}} = E_{\text{complex}} - E_{\text{monomer1}} - E_{\text{monomer2}}$),³² and the associated basis set superposition error (BSSE) was eliminated by the standard counterpoise method of Boys and Bernardi.³³ The ideal and real intermolecular Coulombic energies ($\Delta E_{\text{coul}}^{\text{ideal}}$ and $\Delta E_{\text{coul}}^{\text{real}}$) were calculated in terms of the classical Coulomb's law using the natural charges derived from NBO analysis of the complex members in isolated and in complexed states, respectively. Since Dunning's basis set series is unavailable for iodine, the Lan12DZ basis set, augmented by a set of *d* and *f* polarization functions (exponents 0.292 and 0.441, respectively) and *s* and *p* diffuse functions (exponents 0.0569 and 0.0330, respectively), abbr. Lan12DZ+(df), was used for I[−]. This large version of a valence electron orbit seems to be necessary for reliably describing the outer electronic structure of diffuse anions, and previous theoretical calculations which used this modified effective core potential (ECP) basis set have been shown to give reasonably good results for the I[−]-participating S_N2 reactions³⁴ and the

OCS⋯I[−] van der Waals complexes.³⁵ Because no physical meaning can be ascribed to regions of Cartesian space delimited by zero-flux surfaces derived from the valence electron densities,³⁶ the wave functions generated from the all electron basis set DGDZVP, but not the valence basis set Lan12DZ+(df), were used for AIM analysis of iodine-containing systems.³⁷

To inspect the interaction profile of halogen ions with the protein moieties of interest, a thorough search for all the low-lying energy structures of Cl[−] binding to the electrophilic hydrogen atoms of six protein groups, respectively, modeled by methanol (CH₃OH) (for hydroxyl group), *N*-methylacetamide (CH₃CONHCH₃) (for main chain's amide), acetamide (CH₃CONH₂) (for side chain's amide), methylammonium (CH₃NH₃⁺) (for lysine's ammonium), 4-methylimidazolium



(for histidine's imidazolium), and *N*-methylguanidinium



(for arginine's guanidinium), has been done at the MP2/aug-cc-pVDZ level. No symmetries were constrained in optimization procedures, and the stability of optimized structures was confirmed in the following vibrational frequency analysis.

A two-layer ONIOM-based QM/MM scheme³⁸ was adopted to fully optimize and energetically analyze the protein–ligand interactions through halogen-ionic bridges. The central halogen ion and the corresponding protein residues and ligand that are directly bound to the halogen ion were included in the QM layer and treated at a high level of density functional theory (B3LYP/6-31+G*), while the rest atoms were in the MM layer and treated at a low level of molecular force field (AMBER parm96).³⁹ In the MM layer, water molecules were described by the TIP3P model,⁴⁰ and the restricted electrostatic potential (RESP) fitting procedure⁴¹ was employed to assign partial atomic charges for small ligands and nonstandard amino acid atoms. Parameters that were not found in standard AMBER force fields were defined using the generalized amber force field (GAFF).⁴² Recently, we have successfully employed this ONIOM protocol (only slight modification) to investigate the halogen–water–hydrogen bridges⁴³ and the fluorine bonds⁴⁴ in protein structures and, therefore, expect that this hybrid QM/MM methodology could be used for the halogen-ionic bridge systems as well.

Structure optimizations, energy evaluations, and ONIOM calculations were carried out with the help of the GAUSSIAN 03 suite of programs.⁴⁵ AIM and NBO analyses were implemented in AIM2000⁴⁶ and NBO5.0,⁴⁷ respectively.

2.2. Database Survey. Pretreatment of PDB Files. Up to January, 2010, there were 3391 protein records and 133 nucleic acid entries (solved at 3 Å or better) deposited in the PDB in which at least one nonbonded halogen ion is contained. These structures were extracted and treated with following procedure: (i) removing water molecules, metal

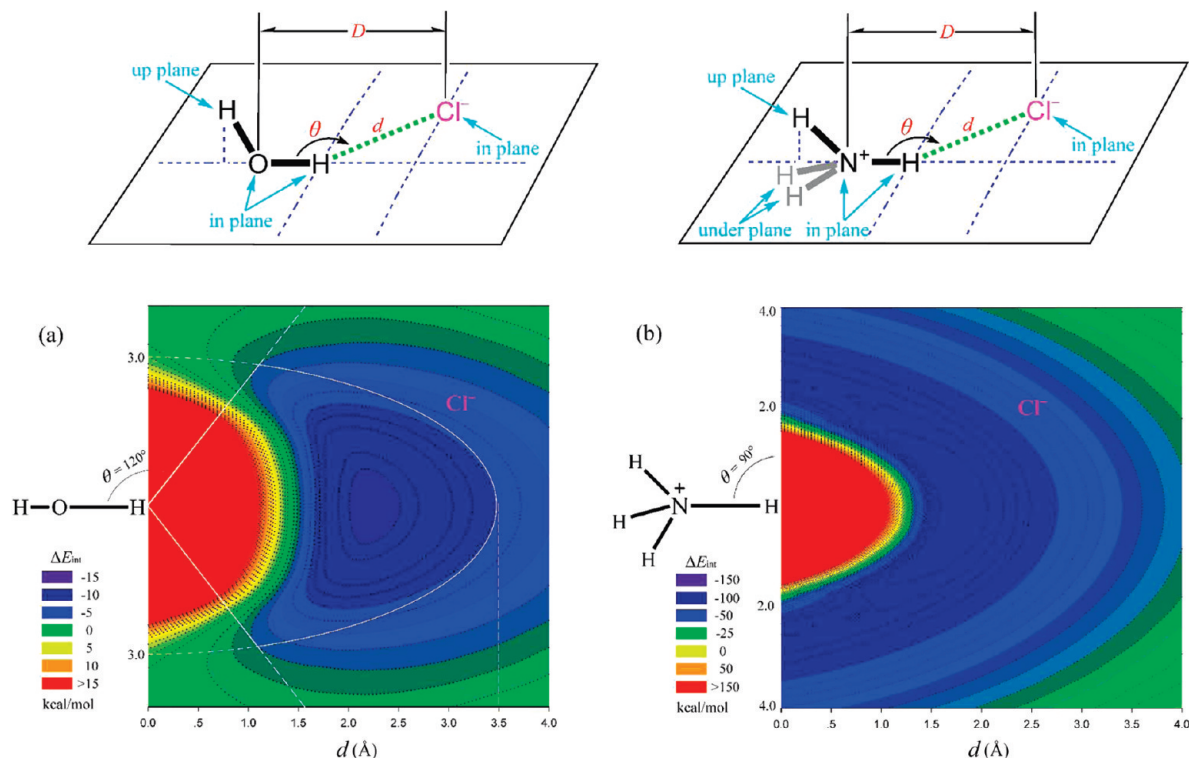


Figure 2. Potential energy surface scans for the complexes $\text{OH}_2 \cdots \text{Cl}^-$ (a) and $\text{NH}_4^+ \cdots \text{Cl}^-$ (b) at the MP2/aug-cc-pVDZ level of theory. In this procedure, the positions of H_2O and NH_4^+ with standard geometries were fixed, and then the probe Cl^- was employed to detect its interaction potentials with the fixed H_2O and NH_4^+ when it ranges over the plane space as shown in top schematic diagrams. As a result, two two-dimensional contours intuitively displaying the intermolecular potential ΔE_{int} as a function of the geometrical arrangement of the complexes were presented. The step sizes for d and θ in the scans were set to 0.2 Å (from 1.0 to 4.0 Å) and 10° (from 90° to 180°), respectively.

ions, and other cofactors, except halogen ions and small organic ligands, (ii) repairing missing side chains of protein residues, using the newly released SCWRL4 program,⁴⁸ (iii) assigning secondary structure class for protein residues, according to the dictionary secondary structure of proteins (DSSP) protocol,⁴⁹ (iv) adding hydrogen atoms for all protein and nucleic acid heavy atoms, using the REDUCE program⁵⁰ (REDUCE was adopted here because this program was tested in our previous study to be capable of precisely reproducing the neutron diffraction-determined hydrogen's positions),⁵¹ (v) defining protonation state of all charged residues at pH 7.0, using the PROPKA 2.0 program,⁵² and (vi) interpreting structural information of small ligands, which are marked by header 'HETATM' in the PDB files, using the I-INTERPRET program.⁵³ This program reads an assembly of ligands in standard PDB format and writes a MOL2 file in which the atomic states, connection manners, and neutral/charged hydrogen's positions are assigned in a considerable accuracy for these ligands.

Definition of Geometrical Constraints. In order to determine the appropriate geometrical constraints used for screening effective interactions between halogen ions and biomolecules in these treated PDB structures, the biomolecular groups that perform as potential halogen ion-acceptors were roughly classified into two types as polar and charged, which were respectively modeled using the H_2O and NH_4^+ , and then, potential energy surface scans for the complexes $\text{OH}_2 \cdots \text{Cl}^-$ and $\text{NH}_4^+ \cdots \text{Cl}^-$ with systematically varying in distance $d_{\text{H} \cdots \text{Cl}^-}$ and angle $\theta_{\angle(\text{O/N}-\text{H} \cdots \text{Cl}^-)}$ have been done at

the MP2/aug-cc-pVDZ level. As a result, two two-dimensional contours intuitively displaying the intermolecular potential ΔE_{int} as a function of the geometrical arrangement of the complexes were presented (Figure 2). A significant difference between these two potential landscapes can be seen. For the complex $\text{NH}_4^+ \cdots \text{Cl}^-$, a strong repulsion appears at the region (red) nearby the interacting H atom, which is surrounded by a prominently attractive area (navy blue) and, farther out, a weak interaction domain (green) (Figure 2b); for the $\text{OH}_2 \cdots \text{Cl}^-$, however, the intermolecular potential is anisotropically distributed around the hydroge (H) atom, a strong attractive potential in the "head on" orientation (navy blue) and a weak repulsive force in the "side on" direction (green) (Figure 2a). According to this finding, together with the conclusions arisen from our other investigations, the following criteria were defined to describe the effective biological interactions involving halogen ions: (i) For a uncharged polar group, an ellipsoid with its center at the polar H atom and its semi-minor/semi-major axis of 3.0/3.5 Å was constructed. Only those halogen ions occurring within the ellipsoidal space and with the forming angle $\theta > 120^\circ$ were considered (i.e., the region encompassed by white solid line in Figure 2a); and (ii) For a charged basic group, the halogen ions with their distances, D , to any one of the heavy atoms in the group less than 4.5 Å were considered. In this way, a halogen-ionic bridge can be readily defined as the entity in which a halogen ion effectively interacts with two or more biomolecular groups simultaneously; the number

Table 1. Experimentally Measured Values of Hydration Enthalpy $\Delta H_{\text{hydr}}^{\circ}$, Hydration Entropy $\Delta S_{\text{hydr}}^{\circ}$, and Hydration Free Energy $\Delta G_{\text{hydr}}^{\circ}$ for Halogen Ions ($T = 298.15$ K)

halogen ion	$\Delta G_{\text{hydr}}^{\circ}$ (kcal·mol ⁻¹) ^a	$\Delta H_{\text{hydr}}^{\circ}$ (kcal·mol ⁻¹) ^a	$-T\Delta S_{\text{hydr}}^{\circ}$ (kcal·mol ⁻¹) ^b
F ⁻	-101.9	-111.5	9.6
Cl ⁻	-73.9	-79.5	5.6
Br ⁻	-70.6	-76.1	5.5
I ⁻	-59.5	-62.3	2.8

^a From ref 59. ^b $-T\Delta S_{\text{hydr}}^{\circ}$ is obtained by subtracting $\Delta H_{\text{hydr}}^{\circ}$ from $\Delta G_{\text{hydr}}^{\circ}$.

of the groups participating in bridging was called the branch degree of this halogen-ionic bridge.

It is worth noting that, although the geometrical criteria presented here were derived on the basis of chlorine ion (the most abundant halogen ion found in biomolecules), this conclusion is also applicable for three other halogen ions.

2.3. Structural and Energetic Properties of Halogen Ions in Halogen-Ionic Bridges. Solvent accessible surface area ($SASA_{\text{brd}}$) and packing density (PD_{brd}) of the bridging halogen ions in protein context can be solved numerically using the, respectively, Sanner's and Voronoi Cell algorithms implemented in the MSMS program⁵⁴ and VORONOA python package⁵⁵ with the ProtOr radii (for protein atoms),⁵⁶ Shannon effective ionic radii (for halogen ions),⁵⁷ and 1.4 Å radii (for water probe). Furthermore, the changes in an ion's hydration enthalpy ($\Delta\Delta H_{\text{hydr}}^{\circ}$), hydration entropy ($\Delta\Delta S_{\text{hydr}}^{\circ}$), and hydration free energy ($\Delta\Delta G_{\text{hydr}}^{\circ}$) due to it transfers from solvent (water) to protein interior and fixed in a halogen-ionic bridge were estimated using the additive models of Ooi et al.:⁵⁸

$$\Delta\Delta H_{\text{hydr}}^{\circ} = \Delta H_{\text{hydr}}^{\circ} \cdot \left(\frac{SASA_{\text{brd}} - SASA_{\text{free}}}{SASA_{\text{free}}} \right) \quad (1)$$

$$\Delta\Delta S_{\text{hydr}}^{\circ} = \Delta S_{\text{hydr}}^{\circ} \cdot \left(\frac{SASA_{\text{brd}} - SASA_{\text{free}}}{SASA_{\text{free}}} \right) \quad (2)$$

$$\Delta\Delta G_{\text{hydr}}^{\circ} = \Delta G_{\text{hydr}}^{\circ} \cdot \left(\frac{SASA_{\text{brd}} - SASA_{\text{free}}}{SASA_{\text{free}}} \right) \quad (3)$$

where $SASA_{\text{brd}}$ and $SASA_{\text{free}}$ are the SASA of the studied halogen ion in bridging and free states, respectively, and $\Delta H_{\text{hydr}}^{\circ}$, $\Delta S_{\text{hydr}}^{\circ}$, and $\Delta G_{\text{hydr}}^{\circ}$ are the standard hydration enthalpy, hydration entropy, and hydration free energy of the halogen ion, i.e., the changes in enthalpy, entropy, and free energy when it moves from the gas phase to a solvent at the standard conditions (1 atm and 298.15 K). The experimentally measured values of $\Delta H_{\text{hydr}}^{\circ}$, $\Delta S_{\text{hydr}}^{\circ}$, and $\Delta G_{\text{hydr}}^{\circ}$ for the four kinds of halogen ions are compiled in Table 1.

2.4. Continuum Electrostatic Analysis. The electrostatic contribution of halogen-ionic bridges to protein stability was ascertained via continuum electrostatic analysis by solving Poisson–Boltzmann (PB) equation, which was implemented in the DELPHI program⁶⁰ with probe radii 1.4 Å, temperature 298.15 K, ionic strength 0.145 M, and dielectric constants 4 for protein and 80 for solvent. A grid spacing of 0.833 Å per grid, in which the longest linear dimension of

the protein occupied 60% of the lattice, was used to determine the size of the cubic lattice, and the Debye–Hückel (full Coulombic) boundary conditions were applied. The PARSE set⁶¹ of partial atomic charges and atomic radii was used for protein atoms, and the formal charge -1 and Shannon effective ionic radii (F⁻ 1.33, Cl⁻ 1.81, Br⁻ 1.96, and I⁻ 2.20 Å)⁵⁷ were assigned for halogen ions.

The total electrostatic contribution ($\Delta\Delta G_{\text{tot}}$) to a halogen-ionic bridge's stability was decomposed into three terms: (i) Bridging energy ($\Delta\Delta G_{\text{brd}}$), which arises from the Coulombic interactions between the halogen-ionic bridge's members (including central halogen ion and its interacting groups) in the folded state of the protein. The $\Delta\Delta G_{\text{brd}}$ can be further divided into two parts: $\Delta\Delta G_{\text{brd}}^{\text{grp}}$, the generally repulsive interaction energy between protein groups in the bridge, and $\Delta\Delta G_{\text{brd}}^{\text{hal}}$, the always attractive interaction energy between the protein groups and halogen ion. (ii) Desolvation cost ($\Delta\Delta G_{\text{dslv}}$), which represents the desolvation penalties incurred by the halogen ion and its interacting partners transferring from a high-dielectric water solvent in the unfolded state to the low-dielectric protein interior in the folded state of the protein. $\Delta\Delta G_{\text{dslv}}$ can also be regarded as the sum of two aspects: $\Delta\Delta G_{\text{dslv}}^{\text{hal}}$ and $\Delta\Delta G_{\text{dslv}}^{\text{grp}}$, the desolvation energies of halogen ions and protein groups, respectively. (iii) Additional effect ($\Delta\Delta G_{\text{add}}$), which accounts for all the Coulombic interactions of the studied halogen-ionic bridge with the charges in rest of the protein in the folded state of the protein. Similarly, $\Delta\Delta G_{\text{add}}$ is broken down into $\Delta\Delta G_{\text{add}}^{\text{hal}}$ and $\Delta\Delta G_{\text{add}}^{\text{grp}}$. These three terms can be readily computed using a strategy proposed by Hendsch and Tidor,⁶² who, and later Kumar et al.,⁶³ had successfully applied this method to investigate protein salt bridges. Briefly, a thermodynamic cycle was performed to trace the changes in Coulombic and reaction field energies of halogen-ionic bridge's members upon the bridge formation during protein folding. In this procedure, electrostatic contribution to free energy change was calculated relative to a mutation of its members to their hydrophobic isosteres. The hydrophobic isosteres were identical with those in the halogen-ionic bridge, except that their partial atomic charges were set to 0. A detailed description of this procedure can be found in refs 62 and 63. The protein moieties, which were considered in the continuum electrostatic calculation as well as their PARSE parameters,⁶¹ employed in this calculation are provided in Supporting Information, Figure S1.

3. Results and Discussion

3.1. Small Model System. I. Electrostatic Potentials. Electrostatic potentials (ESPs),⁶⁴ the most intuitive physical quantity characterizing an electronic distribution state around a molecule or ion, are mapped on the same electronic isodensity surfaces of Cl⁻, H₂O, and NH₃, as shown in Figure 3. The latter two have been widely used as the model of lone-pair donors to study geometrical and energetic features of canonical hydrogen bonding. As might be anticipated, the Cl⁻ performs the typical behavior of hard base with a small radii and a high charge density, which should act as a strong Brønsted base to accept protons donated from the hydrogen-

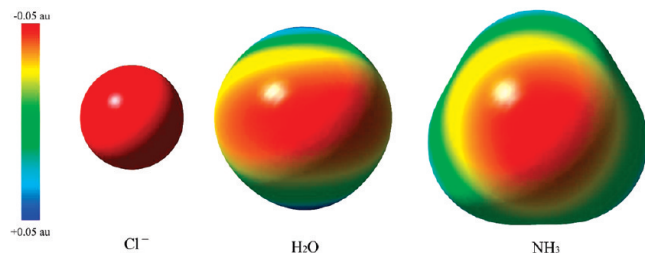


Figure 3. Electrostatic potential distribution, in Hartrees, at the 0.0004 electrons Bohr⁻³ isodensity surfaces of Cl⁻, H₂O, and NH₃.

bond donors, such as serine hydroxyl and lysine ammonium. By contrast, these conventional Lewis bases, H₂O and NH₃, are apparently more ‘soft’ than the Cl⁻, given by the less negative electrostatic potential at their lone-pair sites. In this respect, the Cl⁻, and also the other three halogen ions, expectantly can form strong (ionic) hydrogen bonds, as compared to covalently bonded O and N atoms and can even pair to positively charged biomolecular moieties through ionic bonds.

3.2. Small Model System. II. Intermolecular Interactions. To quantitatively characterize the interaction profile of different halogen ions with polar and charged hydrogen atoms, Figure 4 shows the intermolecular potentials and the Coulombic energies for the complex model systems of H₂O and NH₄⁺ with F⁻, Cl⁻, Br⁻, and I⁻, as a function of the

intermolecular O/N...X⁻ distances. Also plotted are, for comparison purposes, the potential and Coulombic curves of H₂O and NH₄⁺ interacting with the oxygen atom of H₂O, the neutral counterpart of halogen ions. At a first glance, the interactions of halogen ions with H₂O and NH₄⁺ are, as that inferred from ESPs, markedly stronger than that of H₂O lone-pairs with the same hydrogen donors. For example, the well depth of the OH₂...I⁻ potential curve, the most weak complex in the OH₂...X⁻ series, was predicted to be 9.20 kcal·mol⁻¹, which is more than two-fold of the optimal dissociation energy (4.19 kcal·mol⁻¹) of the water dimer (Figure 4a). This difference would increase to six-fold when the hydrogen donor H₂O is replaced by NH₄⁺ (~120 kcal·mol⁻¹ for NH₄⁺...X⁻ vs ~20 kcal·mol⁻¹ for NH₄⁺...OH₂) (Figure 4d). Comparison of intermolecular potentials to ideal and real Coulombic energies for both OH₂...OH₂ (Figure 4b) and OH₂...Cl⁻ (Figure 4c) adducts preliminarily sheds light on the physical nature of these interactions; OH₂...Cl⁻ bonding is dominated by an electrostatic force, given by the good agreement of its potential curve with the corresponding real Coulombic curve. The charge transfers (CTs) seem to be significant in the OH₂...Cl⁻ system, which is implied by the large deviation of the real Coulombic curve from the corresponding ideal one (the former was calculated based on the real charge distribution in the complex system, whereas the latter based on the atomic charges of isolated complex members). In

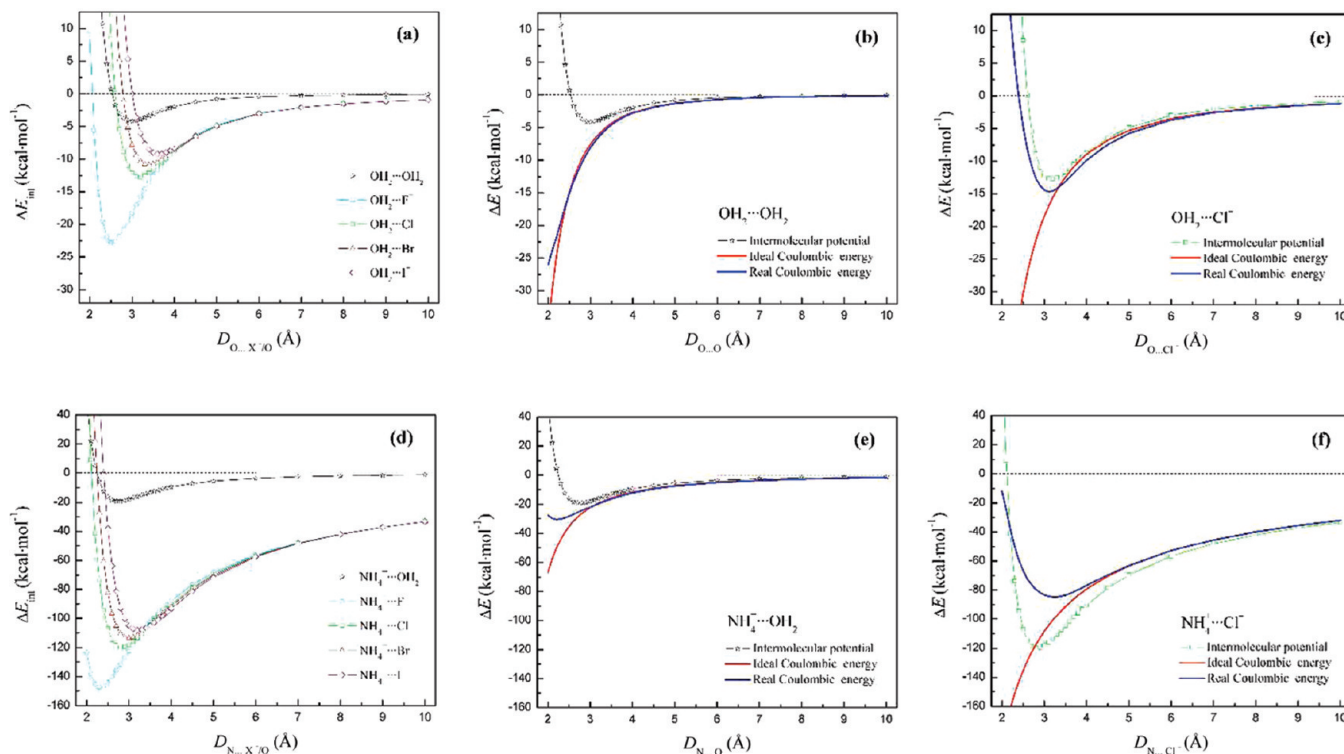


Figure 4. (a) Intermolecular potentials for the complexes of H₂O with F⁻, Cl⁻, Br⁻, I⁻, and H₂O. (b) Comparison of intermolecular potential to ideal and real Coulombic energies for complex OH₂...OH₂. (c) Comparison of intermolecular potential to ideal and real Coulombic energies for complex OH₂...Cl⁻. (d) Intermolecular potentials for the complexes of NH₄⁺ with F⁻, Cl⁻, Br⁻, I⁻, and H₂O. (e) Comparison of intermolecular potential to ideal and real Coulombic energies for complex NH₄⁺...OH₂. (f) Comparison of intermolecular potential to ideal and real Coulombic energies for complex NH₄⁺...Cl⁻. All of energetic data plotted here were determined at the MP2/aug-cc-pVDZ (or MP2/Lan12DZ+(df) for iodine) level of theory. Limited by space, comparisons of intermolecular potentials to Coulombic energies for the complexes of H₂O and NH₄⁺ with F⁻, Br⁻, and I⁻ are provided in the Supporting Information, Figures S2 and S3.

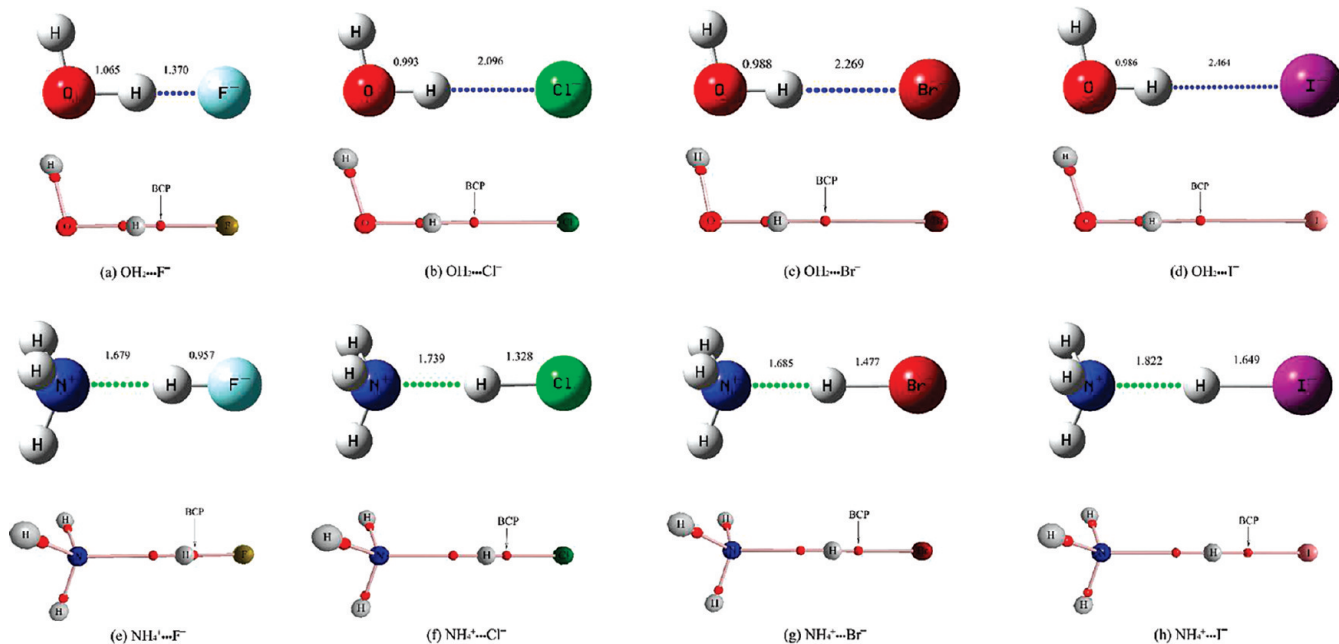


Figure 5. Equilibrium structures and corresponding molecular graphs for complex model systems $\text{OH}_2 \cdots \text{F}^-$ (a), $\text{OH}_2 \cdots \text{Cl}^-$ (b), $\text{OH}_2 \cdots \text{Br}^-$ (c), $\text{OH}_2 \cdots \text{I}^-$ (d), $\text{NH}_4^+ \cdots \text{F}^-$ (e), $\text{NH}_4^+ \cdots \text{Cl}^-$ (f), $\text{NH}_4^+ \cdots \text{Br}^-$ (g), and $\text{NH}_4^+ \cdots \text{I}^-$ (h). In the equilibrium structures, the shorter one of the bonding between O/N and H and between X and H is plotted as a solid line, while the longer one is plotted as a dotted line. Bond lengths are shown in Å. Optimizations were performed at the MP2/aug-cc-pVTZ (or MP2/Lan12DZ+(df) for iodine) level with the angle $\theta_{\angle(\text{O/N}-\text{H} \cdots \text{X}^-)}$ frozen in 180° to avoid secondary interactions between halogen ions and other hydrogen atoms in these systems. Molecular graphs were generated by AIM analysis of corresponding MP2-optimized geometries.

contrast, the real Coulombic curve of the $\text{OH}_2 \cdots \text{OH}_2$ dimer is coincidentally well with the ideal, indicating only a slight CT accompanied with the formation of this complex. These phenomena basically recurred when we investigated the complex systems of $\text{NH}_4^+ \cdots \text{OH}_2$ (Figure 4e) and $\text{NH}_4^+ \cdots \text{Cl}^-$ (Figure 4f), albeit not completely recurred. Most unexpectedly, the real Coulombic curve of the doubly charged $\text{NH}_4^+ \cdots \text{Cl}^-$, unlike that of $\text{OH}_2 \cdots \text{Cl}^-$, is almost on the above of and deviated considerably from the intermolecular potential curve, this phenomenon is quite obvious in the region nearby curve minimum. It means that, besides the electrostatic effect, there also exists other chemical force(s) to contribute the $\text{NH}_4^+ \cdots \text{Cl}^-$ attraction, but what is (are) it? As we know, the $\text{NH}_4^+ \cdots \text{Cl}^-$ complex can also be viewed as formed by NH_3 with HCl , if considering that the proton bound between N^+ and Cl^- is always mobile. In this regard, the $\text{H} \cdots \text{Cl}^-$ interaction (maybe written as $\text{H}^{\delta+} - \text{Cl}^{\delta-}$ bond is more suitable) is, to some extent, imparted with a covalent component and thereby tagged with an additional stabilization energy to enhance its bonding strength. This hypothesis will be proved by topological analysis of electron density and chemical bonding in these model systems (vide post).

3.3. Small Model System. III. Geometrical, Energetic, And Electronic Analyses. The complex model systems of H_2O and NH_4^+ with F^- , Cl^- , Br^- , and I^- were further fully optimized at the MP2/aug-cc-pVTZ (or MP2/Lan12DZ+(df) for iodine) level with the angle $\theta_{\angle(\text{O/N}-\text{H} \cdots \text{X}^-)}$ frozen in 180° to avoid secondary interactions between halogen ions and other hydrogen atoms in these systems. The equilibrium structures and corresponding molecular graphs, which were generated by AIM analysis of the MP2 optimized geometries,

are shown in Figure 5 and parametrized as that listed in Table 2. It is evident that the halogen ions in all complexes have a pronounced interaction with the hydrogen donors, rationalized by the presence of bond paths linking the nuclei of X^- and H. In addition, the interatomic separations between halogen and hydrogen are always longer than corresponding O–H bonds in $\text{OH}_2 \cdots \text{X}^-$, but this is converse in the $\text{NH}_4^+ \cdots \text{X}^-$ series (due to the proton transfers). Specifically, the $\text{H} \cdots \text{F}$ distance in equilibrium $\text{NH}_4^+ \cdots \text{F}^-$ structure is only 0.957 Å, which follows the typical feature of open-shell (shared) interactions. Hence, rather than the nonbonded intermolecular force, it would better be recognized as a polar covalent or ionic bond. Qualitative graphic conclusions could be substantiated by quantitative examination of the geometrical, energetic, and electronic parameters associated with these interactions (Table 2). The first and most straightforward evidence is the particularly high interaction energies (up to $175 \text{ kcal} \cdot \text{mol}^{-1}$) attached to the $\text{NH}_4^+ \cdots \text{X}^-$ associations, these values fall within the normal range of the bond energies of covalent and ionic bonds. By contrast, the intermolecular potentials of $\text{OH}_2 \cdots \text{X}^-$ are only a tenth of that found in corresponding NH_4^+ -involved adducts, satisfying the definition of ionic hydrogen bonds by Meot-Ner.⁶⁵ Second, the unshared attribute for the $\text{OH}_2 \cdots \text{X}^-$ and the shared character for the $\text{NH}_4^+ \cdots \text{X}^-$ can be clearly characterized by the electronic topological parameters (including electron density ρ_b , Laplacian of the electron density $\nabla^2 \rho_b$, ellipticity ϵ_b , and electronic energy density H_b) at the bond critical points (BCPs) of $\text{H} \cdots \text{X}^-$ bond paths (as marked in Figure 5). For example, ρ_b and $\nabla^2 \rho_b$ of $\text{OH}_2 \cdots \text{X}^-$ were predicted to be in the range of 0.006–0.096 and 0.037–0.121 au, respectively, which are basically consistent with or

Table 2. Geometrical, Energetic, and Electronic Parameters for the Complexes of Halogen Ions with H₂O and NH₄⁺ Serving As Hydrogen Donor^a

Complex	$d_{\text{H}\cdots\text{X}}^{\text{b}}$	$D_{\text{Y}\cdots\text{X}}^{\text{c}}$	$\Delta E_{\text{int}}^{\text{d}}$	$\rho_{\text{b}}^{\text{e}}$	AIM analysis		H_{b}^{h}	Wiberg ^f	NLMO/NPA ^g	$\Delta q_{\text{X}}^{\text{k}}$	NBO analysis		
					$\nabla^2 \rho_{\text{b}}^{\text{i}}$	$\varepsilon_{\text{b}}^{\text{g}}$					$\Delta E_{\text{coll}}^{\text{ell}}$	$\Delta E_{\text{SE}}^{\text{m}}$	
OH ₂ ⋯F ⁻	1.370	2.435	-26.61 (-23.32) ^q	0.0958	0.1207	0.0002	-0.0514	0.1699	0.0883	0.1095	-32.02	51.12	LP(2)F ⁻ →BD ^o (1)O-H: 1.78 LP(3)F ⁻ →BD ^o (1)O-H: 4.99 LP(4)F ⁻ →BD ^o (1)O-H: 92.86
OH ₂ ⋯Cl ⁻	2.096	3.089	-14.11 (-14.71) ^r	0.0328	0.0585	0.0014	-0.0062	0.0701	0.0356	0.0458	-16.42	18.58	LP(1)Cl ⁻ →BD ^o (1)O-H: 0.62 LP(4)Cl ⁻ →BD ^o (1)O-H: 23.37
OH ₂ ⋯Br ⁻	2.269	3.257	-11.99 (-11.71) ^r	0.0277	0.0480	0.0012	-0.0041	0.0615	0.0314	0.0404	-14.22	16.56	LP(1)Br ⁻ →BD ^o (1)O-H: 0.44 LP(4)Br ⁻ →BD ^o (1)O-H: 18.95
OH ₂ ⋯I ⁻	2.464	3.449	-9.80 (-10.30) ^r	0.0056	0.0371	0.0008	-0.0036	0.0603	0.0304	0.0398	-11.76	15.72	LP(1)I ⁻ →BD ^o (1)O-H: 0.33 LP(4)I ⁻ →BD ^o (1)O-H: 16.97
NH ₄ ⁺ ⋯F ⁻	0.957	2.636	-175.56	0.3186	-2.5734	0.0002	-0.7423	0.5716	0.3489	0.3600	-119.17	NA ^p	NA ^p
NH ₄ ⁺ ⋯Cl ⁻	1.328	3.067	-131.98	0.2219	-0.7107	0.0001	-0.2302	0.8106	0.6107	0.6135	-26.96	NA ^p	NA ^p
NH ₄ ⁺ ⋯Br ⁻	1.477	3.162	-120.63	0.1808	-0.4494	0.0000	-0.1584	0.8245	0.6512	0.6552	-17.47	NA ^p	NA ^p
NH ₄ ⁺ ⋯I ⁻	1.649	3.471	-113.17	0.1466	-0.1506	0.0000	-0.1113	0.9129	0.8151	0.8198	-2.90	NA ^p	NA ^p

^a The equilibrium structures as well as AIM and NBO parameters of all complexes were obtained at the MP2/aug-cc-pVTZ (or MP2/Lan12DZ+(df) for iodine for structure optimization and NBO analysis or at MP2/DGDZVP for iodine for AIM analysis) level of theory. Interaction energies ΔE_{int} between halogen ions and their interacting partners were calculated at a higher level of CCSD(T)/aug-cc-pVTZ (or CCSD(T)/Lan12DZ+(df) for iodine) using the MP2 optimized geometries. ^b $d_{\text{H}\cdots\text{X}}$ (Å), interatomic distance between X⁻ and H (for OH₂⋯X⁻ and NH₄⁺⋯X⁻), where X⁻ is halogen ion. ^c $D_{\text{Y}\cdots\text{X}}$ (Å), interatomic distance between Y and X⁻, where X⁻ is halogen ion and Y is O or N in H₂O or NH₄⁺, respectively. ^d ΔE_{int} (kcal·mol⁻¹), calculated and, if exist, experimental (in parentheses) interaction energies. ^e ρ_{b} (au), electron density at BCPs. ^f $\nabla^2 \rho_{\text{b}}$ (au), Laplacian of the electron density at BCPs. ^g H_{b} (au), electronic energy density at BCPs. ^h Wiberg, Wiberg bond index associated with the interactions. ⁱ NLMO/NPA, atom-atom net linear NLMO/NPA bond order associated with the interactions. ^k Δq_{X} , changes in charge number of halogen ion due to the interactions. ^l $\Delta E_{\text{coll}}^{\text{ell}}$ (kcal·mol⁻¹), real Coulombic energy between donor and acceptor, obtained via Coulomb's analysis. ^m ΔE_{SE} (kcal·mol⁻¹), steric exchange energy between halogen ion and its interacting partners. ⁿ ΔE^{z} (kcal·mol⁻¹), second-order perturbation stabilization energy between the lone pair of halogen ion and the antibonding orbital σ^* of O-H or N-H bond in H₂O or NH₄⁺, respectively. ^o Not applicable. ^p From gas-phase equilibrium measurements by high-pressure mass spectrometry. ^q From gas-phase equilibrium measurements by pulsed electron beam high-pressure mass spectrometry. ^r From gas-phase equilibrium measurements by high-pressure mass spectrometry.

slightly larger than that proposed for (nonionic) hydrogen bonds (i.e., 0.002–0.035 and 0.024–0.139 au, respectively),⁶⁶ but these quantities are noticeable in the $\text{NH}_4^+\cdots\text{X}^-$ repertoire. In addition, the negative values of H_b for the both $\text{OH}_2\cdots\text{X}^-$ and $\text{NH}_4^+\cdots\text{X}^-$ suggest that these two kinds of interactions are stronger than conventional hydrogen bonding, which usually has a positive H_b .⁶⁶ The last but most importantly, NBO population analysis of the natural atomic and bond orbits in these complexes depicted a clearer profile for the chemical nature of interactions involving halogen ions. As listed in Table 2, the Wiberg⁶⁷ and natural localized molecular orbital/natural population analysis (NLMO/NPA)⁶⁸ bond indices of the $\text{NH}_4^+\cdots\text{X}^-$ interactions are quite significant with respect to Cl^- , Br^- , and, particularly, I^- (close or greater than 0.6) but relatively lower in F^- -containing system. This is inconsistent with that reflected in interaction energies ΔE_{int} , which enhances in the order of $\text{NH}_4^+\cdots\text{I}^- < \text{NH}_4^+\cdots\text{Br}^- < \text{NH}_4^+\cdots\text{Cl}^- < \text{NH}_4^+\cdots\text{F}^-$. It must be recalled here that most of bond orders (BOs), such as the Wiberg and NLMO/NPA discussed here, generally underestimate for polar covalent bonds, since the ionic bond component in these polar covalent bonds is almost ignored by BOs. Compared to those found in $\text{NH}_4^+\cdots\text{Br}^-/\text{Cl}^-/\text{I}^-$ complexes, the relatively small degree of charge transfers ($\Delta q_{\text{F}^-} = 0.36$) and the dominant Coulombic effect ($\Delta E_{\text{col}}^{\text{real}} = -119.17 \text{ kcal}\cdot\text{mol}^{-1}$) unravel a marked ionic bond character associated with the $\text{NH}_4^+\cdots\text{F}^-$ interaction, because these electrostatic properties are always related to the ionic bonding.⁶⁹ Furthermore, it must be reminded here that complicated biological context would undermine the “idea fashion” (as that in small model systems) of halogen ions approaching charged hydrogen atoms, giving rise to a more important role of the nondirectional ionic bonding than the directional covalent bonding in the interactions involving not only F^- but also other three halogen ions.

Overall, halogen ions in physiological environment are presumed to adopt three types of interactions to bridge between biomolecular moieties: (i) ionic hydrogen bonding with polar hydrogen atoms, such as those in amide and hydroxyl group; (ii) ionic bonding with positively charged species, such as ammonium, imidazolium, and guanidinium; and (iii) covalent bonding with the hydrogen atoms of proton donors (this interaction type must have a “good” geometrical arrangement as compared to ionic bonding). It should be noted here that this division is not absolute and most of the real cases must be compatible simultaneously with two or even three of these interaction types.

3.4. Real Biomolecular System. I. PDB Survey of Halogen-Ionic Bonding. The PDB (January, 2010 release) contains 3391 and 133 entries of X-ray crystal structures (at resolutions of 3.0 Å or better) of proteins and nucleic acids showing 11 852 and 1345 nonbonded halogen ions, respectively (total 37 F^- , 9966 Cl^- , 1065 Br^- , and 2129 I^-). The pronouncedly unbalanced numbers of different halogen ions found in biomolecules mirror the variance in chemical activity and the natural abundance of these halogens. Fluorine is the most active element in halogens and usually exists in combined state. Hence, only a few of free fluorine ions were observed in the survey. However, the also chemically active

chlorine ions were found to be quite abundant in biological systems. This could be attributed to the important physiological function of the chlorine ions in keeping, for example, electrical neutrality, acid–base balance, and correct pressure of cell and body. Using the criteria described in Section 2.2, we selected the halogen ions in effective interaction (for convenience, termed as halogen-ionic bonding) with the polar hydrogen atoms and positively charged groups of biomolecules to define a reliable set of solid biological contacts involving halogen ions, which consists of 20 826, 793, and 174 halogen-ionic bonding with proteins, nucleic acids, and small ligands, respectively. Considering the prominent magnitude of halogen-ionic bonding with protein moieties, we herein give a detailed inspection on the geometrical characteristics and distribution of this kind of interactions.

Relative halophilicity of different protein moieties was assessed using their contact rates (CRs) with halogen ions, which were simply defined as the ratio of those in effective interaction with halogen ions to all presented in our data set. As might be expected, three charged moieties, i.e. ammonium, imidazolium, and guanidinium, have the highest propensity to pair with halogen ions, with their CRs of 2.34, 4.32, and 5.47%, respectively. In the remaining polar moieties, the side chain’s amide performs as well in the halophilicity (CR = 1.62%), whereas the main chain’s amide and hydroxyl group exhibit a relatively halophobic feature (CRs < 1%). The difference in halophilicity between polar and charged protein moieties well echoes the bonding strength variation among small model systems and also reflects the electrostatic nature of biological halogen-ionic bonding. These conclusions could be further visualized by superposing halogen ions around their common protein moieties and by comparing the distribution states of these superposed halogen ions to corresponding MP2-determined low-lying energy structures, as shown in Figure 6. It can be seen that the main chain’s amide and hydroxyl group hold only one hydrogen site to accommodate halogen ions, hence, their halophilicity is lower than the congeneric side chain’s amide, which provides two hydrogen sites for halogen ions. The polyatomic imidazolium and guanidinium have a large surface to contact with surrounding halogen ions and thus show the highest halophilicity, while the smaller ammonium can only possess a moderately strong halophilicity. Besides, the distribution preferences of halogen ions around different protein moieties are compatible with the Cl^- locations in corresponding low-lying energy structures. For example, the arrangement patterns of Cl^- in interaction with three charged moieties in low-lying energy structures are clearly mirrored as the presence of local halogen ion-dense regions in the statistical distribution plots, although the contacting angle of halogen ions with these charged moieties has been entirely ignored when we performed the PDB survey.

Despite the significant unbalancedness of four kinds of halogen ions occurring in a biological environment, the contacting behavior of different halogen ions with protein moieties could also be analyzed in terms of the frequency distributions of geometrical parameters of halogen-ionic bonds (HIBs) retrieved from the PDB. It is shown that the geometrical profile of polar HIBs (bonding between halogen

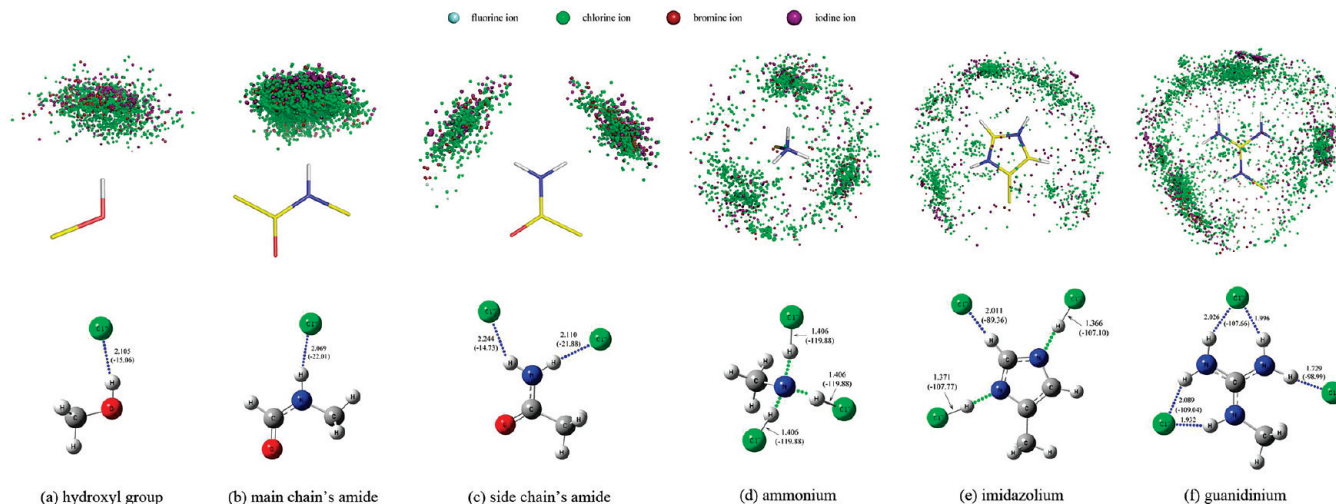


Figure 6. The first row showing the distribution states of halogen ions around different protein moieties retrieved from the PDB. The second row showing the low-lying energy structures of Cl^- in complex with corresponding protein moieties obtained by a thorough MP2/aug-cc-pVDZ search (to render this figure more readable, multiple low-lying energy sites of Cl^- in complex with the same moiety are artificially resettled in a subplot).

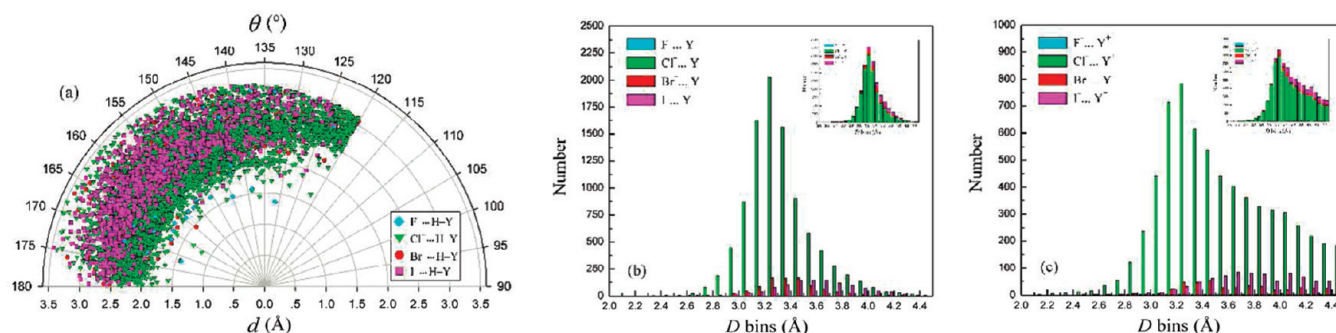


Figure 7. (a) Polar scatter plot of θ vs d for halogen-ionic bonding with polar hydrogen atoms of proteins retrieved from the PDB. Histogram distributions of D (in 0.1 bins) for halogen-ionic bonding with: (b) polar and (c) charged moieties of proteins retrieved from the PDB. θ is the angle $\angle(\text{O}-\text{H} \cdots \text{X}^-)$ for hydroxyl groups or the angle $\angle(\text{N}-\text{H} \cdots \text{X}^-)$ for amides; d is the interatomic distance between X and polar H in polar moieties; D is the interatomic distance between X and heavy atom Y , where Y is the antecedent of the interacting polar H in polar moieties or the nearest heavy atom in charged moieties.

ions and polar protein groups) is in agreement with those received from statistical examination of water–halide ion interactions in the CSD⁷⁰ and routine hydrogen bonds in protein crystals,⁷¹ albeit the d values of HIBs seem to be slightly larger than that of hydrogen bonds (Figure 7a). This is not unexpected if considering that these polar HIBs are natural of ionic hydrogen bonds, and the longer interatomic separations between H and X^- in polar HIBs than those between H and O/N in routine hydrogen bonds are apparently owing to the larger radii of halogen ions (except fluorine ion) relative to oxygen and nitrogen atoms. Moreover, from the frequency distributions of bond lengths D derived from polar and charged HIBs (Figure 7bc), it can be readily appreciated that the D values also increase with the size of halogen ions, i.e. $\text{F}^- \cdots \text{Y}/\text{Y}^+ < \text{Cl}^- \cdots \text{Y}/\text{Y}^+ < \text{Br}^- \cdots \text{Y}/\text{Y}^+ < \text{I}^- \cdots \text{Y}/\text{Y}^+$. Interestingly, the peak positions of D distributions for three polar HIBs ($\text{Cl}^- \cdots \text{Y}$, $\text{Br}^- \cdots \text{Y}$, and $\text{I}^- \cdots \text{Y}$) and for three charged HIBs ($\text{Cl}^- \cdots \text{Y}^+$, $\text{Br}^- \cdots \text{Y}^+$, and $\text{I}^- \cdots \text{Y}^+$) are completely consistent, as both series located at the 3.25, 3.45, 3.65 Å bins, respectively ($\text{F}^- \cdots \text{Y}$ and $\text{F}^- \cdots \text{Y}^+$ were not considered here because their numbers found in the PDB are too small to generate

statistically significant conclusions), and these peak locations are uniformly accompanied with a red-shift of about 0.2 Å relative to equilibrium distances in corresponding small model complexes (see Table 2). The elongating of HIBs in biomolecules could be ascribed to steric hindrance and constraint in complicated biological context. Furthermore, although polar and charged HIBs have a coherency in their D peak locations, the whole profile of their D distributions is solidly distinct, particularly in the regions to the right of the peak positions (Figure 7b and c). Compared to polar HIBs, charged are more long-range and hold a considerable number with bond lengths $D > 4.0$ Å. This can be reflected in the intermolecular potential curves derived from small model systems (Figure 4a and d). At the 4.5 Å separation, for example, intermolecular potentials for $\text{OH}_2 \cdots \text{X}^-$ models are only about $-7 \text{ kcal} \cdot \text{mol}^{-1}$, while for $\text{NH}_4^+ \cdots \text{X}^-$ models are of striking values as more than $-80 \text{ kcal} \cdot \text{mol}^{-1}$. The fundamental difference between polar and charged HIBs in long-range interaction behavior reveals and substantiates their natures of ionic hydrogen bonding and ionic bonding, respectively.

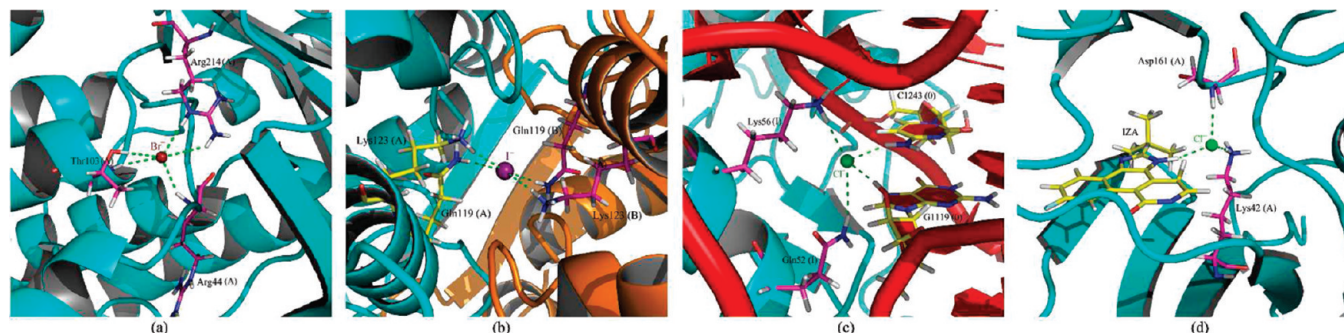


Figure 8. Some examples of halogen-ionic bridges in biomolecules. (a) Bromine-ionic bridge in protein interior (PDB: 1doc). (b) Iodine-ionic bridge at protein–protein interface (PDB: 2vgz). (c) Chlorine-ionic bridge at protein–rRNA interface (PDB: 1jj2). (d) Chlorine-ionic bridge at protein–small ligand interface (PDB: 2j90).

3.5. Real Biomolecular System. II. PDB Survey of Halogen-Ionic Bridges. On the basis of the HIBs retrieved from the PDB, we have defined a distinct data set consisting of 6406 biological halogen-ionic bridges; each of them is composed of a halogen ion and several biomolecular moieties that are directly bound to the halogen ion through HIBs. This substantial magnitude of halogen-ionic bridges found here is remarkably more than the numbers of halogen bonds (113)⁷² and halogen–water–hydrogen bridges (106)⁴³ surveyed in previous works, implying a non-negligible role of the halogen-ionic bridges in biological systems. We further classified these halogen-ionic bridges into four groups according to their locations, i.e., 4910 in protein interiors, 1132 at protein–protein interfaces, 273 at protein–nucleic acid interfaces, and 91 at protein–small ligand interfaces. A classification list of these halogen-ionic bridges is tabulated in Supporting Information, Tables S1–S4.

Halogen-Ionic Bridges in Protein Interiors (Figure 8a). A considerable number of halogen-ionic bridges stand in protein interiors and are thought to be functional in figuring the advanced structure of protein architectures. Formation of halogen-ionic bridges in the low dielectric environment due to protein folding causes a large desolvation penalty, which can be, more or less, compensated by the favorable electrostatic interactions both between the halogen ions and their oppositely charged partners and between the halogen-ionic bridges and their protein surroundings. In the next section, we will provide computational evidence supporting that halogen-ionic bridges are generally stabilizing toward protein architectures, and this stabilization tendency is quite significant as compared to the diverse salt bridges described previously. In addition, most of the halogen-ionic bridges in proteins are formed between sequentially farther residues in comparison with salt bridges, which usually pair within the vicinal residue blocks.⁶³ The average number of amino acids separating the halogen-ionic bridging residues is 24.01, this is far beyond that to be considered in the hierarchical model of protein folding.⁷³ Hence, halogen-ionic bridges in proteins are most likely to be formed later than the “molten globule” phase of the folding.

Halogen-Ionic Bridges at Protein–Protein Interfaces (Figure 8b). Owing to the chemically natural similarity between protein–protein interface and protein interior,⁷⁴ halogen-ionic bridges at the interfaces are supposed to confer stability and specificity for protein binding as much as that for protein

folding. In fact, halogen-ionic bridges seem to be more effective in contributing to protein binding rather than to folding because the binding does not need too much of a degree of packing the halogen-ionic bridging groups from the already structured protein monomers to complex, thus leading to a lesser desolvation penalty.

Halogen-Ionic Bridges at Protein–Nucleic Acid Interfaces (Figure 8c). Almost all of the halogen-ionic bridges across protein–nucleic acid interfaces were found in huge, complicated ribosomes. Since ribosome growth is an exhaustive process which involves a series of molecular operation steps, such as protein/rRNA splicing, folding, and packing,⁷⁵ the role played by halogen-ionic bridges in the ribosomes should be different to those in protein complexes, which are normally formed by direct, rigid protein–protein association. Visual inspection of these ribosomal halogen-ionic bridges found that they are usually located at the regions where protein and rRNA atoms are fully buried but not in sufficient contact with each other, the halogen ions occupy at the cavities embedded within these atoms and interact with vicinal polar groups. On this point, the halogen-ionic bridges at protein–nucleic acid (rRNA) interfaces could be regarded as structural fillers to refine the shape complementarity and to tune the local conformation of the interface structures.

Halogen-Ionic Bridges at Protein–Small Ligand Interfaces (Figure 8d). Although limited numbers of halogen-ionic bridges were observed at the protein–small ligand interfaces, we will demonstrate that they do play an important role in inhibitor recognition and binding, at least by HIV-1 protease and PDK1 kinase, using the QM/MM scheme (see Section 3.7). Statistical analysis unraveled that halogen-ionic bridges at protein–ligand interfaces share a similar solvent accessibility (measured by SASA) and fluctuation rate (measured by isotropic *B*-factor) with those ligand-bound water molecules,⁷⁶ but they usually stand in multifurcated form (measured by branch degree) and are packed tightly by surrounding protein and ligand atoms (measured by packing density). In general, the halogen-ionic bridge, if it exists, is shown to be essential in assisting the ligand positioning in the protein active pocket because only the accurate ligand location can result in the bridge with optimal geometry and thus highest stability (this will be rationalized by QM/MM procedure).

The mean statistics of structural and energetic parameters for the 6042 halogen-ionic bridges found in proteins (includ-

ing protein interiors and interfaces) are compiled in Table 3. Nearly 35% of halogen-ionic bridging amino acids are located at a protein helix region, and the remainders (~65%) are approximately equivalently distributed in strand, turn, and loop. This assignment agrees to the abundance of these secondary structure classes observed in native proteins.⁷⁷ Comparison of the average *B*-factors between bridging and nonbridging halogen ions suggested that halogen ions are generally less mobile when they are bound in halogen-ion bridges than when they are out of the bridges, given by ~20% difference in their average *B*-factor values. This phenomenon has also been observed for the water-mediated bridges in protein crystals.⁷⁶ In addition, the fluorine-ionic bridge has a large branch degree and packing density as compared to the other three; this could be reflected in the significant desolvation effect accompanied with the fluorine-ionic bridge formation. As can be seen in Table 3, the average percentage of reduction in the SASA when the F⁻ transfer from solvent to bridges is 87.7%; this, coupled with its prominent thermodynamic effect of hydration, gives rise to the F⁻ with a noticeable value in $\Delta\Delta G^{\circ}_{\text{hydr}}$, $\Delta\Delta H^{\circ}_{\text{hydr}}$, and $-T\Delta\Delta S^{\circ}_{\text{hydr}}$. In contrast, the desolvation penalties of bridging Cl⁻, Br⁻, and I⁻ have only about two-thirds of that with the bridging F⁻. In this regard, the desolvation profile of formation of different halogen-ionic bridges is compatible with the specific ion effects observed experimentally, i.e., F⁻ is referred as kosmotrope, whereas Cl⁻, Br⁻, and I⁻ are called chaotropes.³

3.6. Real Biomolecular System. III. Continuum Electrostatic Analysis. Based on conventional biochemical intuition, one would expect halogen-ionic bridges to be stabilizing toward the folded conformations of proteins. However, we should be cautious of this notion, recalling that some salt bridges have been demonstrated to be destabilizing for protein structures since their desolvation penalties, due to the burial of ionizable salt-bridging groups in the low dielectric protein interior during protein folding, are not fully recovered by favorable electrostatic interactions in the folded state.^{78–80} So, we herein performed continuum electrostatic calculations on a panel of high-quality halogen-ionic bridges derived from monomeric protein crystal structures to answer questions like do the halogen-ionic bridges generally stabilize protein architectures and whether they confer more stability for proteins than traditional salt bridges? Since all the calculations are essentially based upon the atomic coordinates provided in protein PDB files, the accuracy of the results gained from the continuum electrostatic analysis is highly dependent on the quality of the protein structures in which the studied halogen-ionic bridges are contained. Therefore, we used a set of monomeric protein structures with high resolution (≤ 1.8 Å) and low homology (sequence identity <30% between any two proteins) taken from the current (October 14, 2009 updated) PDB-REPRDB list of structures.⁸¹ From this list, we culled the halogen-ionic bridges satisfying the criteria as defined in Section 2.2, and also which the positions of the central halogen ions possess a high precision (occupancy = 1 and *B*-factor <30 Å²). Consequently, 241 high-quality halogen-ionic bridges distributed in 189 monomeric proteins were selected for

Table 3. Mean Statistics of Structural and Energetic Parameters for Halogen-Ionic Bridges in Proteins^a

ion	num.	SSC ^b	BF _{brid} ^c	BD _{brid} ^d	PD _{brid} ^e	SASA _{brid} ^f	Δ SASA _{brid} % ^g	$\Delta\Delta G^{\circ}_{\text{hydr}}$ ^h	$\Delta\Delta H^{\circ}_{\text{hydr}}$ ⁱ	$-T\Delta\Delta S^{\circ}_{\text{hydr}}$ ^j
F ⁻	18	46.5% H, 13.5% S, 17.7% T, 22.3% L	34.47(13.52)	3.22 (1.48)	0.74 (0.04)	11.49 (12.79)	87.7%	89.40 (13.91)	97.82 (15.22)	-8.42 (1.31)
Cl ⁻	5074	37.1% H, 19.9% S, 21.5% T, 21.4% L	37.63(19.04)	2.56 (0.78)	0.40 (0.12)	20.80 (20.18)	83.9%	62.03 (11.52)	66.73 (12.39)	-4.70 (0.87)
Br ⁻	296	30.8% H, 32.0% S, 23.6% T, 13.7% L	32.87(19.93)	2.41 (0.66)	0.41 (0.11)	25.01 (19.63)	82.4%	58.15 (9.77)	62.69 (10.53)	-4.53 (0.76)
I ⁻	553	30.1% H, 19.8% S, 27.2% T, 22.9% L	46.30(23.52)	2.29 (0.53)	0.40 (0.10)	33.57 (25.99)	79.4%	47.23 (9.50)	49.46 (9.94)	-2.22 (0.45)
All	6042	34.1% H, 23.5% S, 21.3% T, 21.1% L	38.09(19.72)	2.53 (0.76)	0.40 (0.11)	22.22 (21.06)	83.3%	60.50 (12.14)	64.98 (13.19)	-4.47 (1.12)

^a The numbers in parentheses are corresponding standard deviations. ^b SSC, secondary structure class assignment for the amino acids that are directly bound to the bridging halogen ions. H, helix (α -, π -, and 3/10-helix); S, strand (isolated β -strand and multiple β -ladder); T, turn (3, 4, 5 turns and bend); L, loop and others (loop, coil, etc.). ^c BF_{brid} (Å²), mean *B*-factor (or Debye–Waller factor) of the bridging halogen ions. ^d BD_{brid}, mean branch degree of the halogen-ionic bridges. ^e PD_{brid}, mean pack density of the bridging halogen ions. ^f SASA_{brid} (Å²), mean SASA of the bridging halogen ions. ^g Δ SASA_{brid}%, average percentage of changes in the SASA when the halogen ions transfer from solvent (free state) to halogen-ionic bridges (bridging state). ^h $\Delta\Delta G^{\circ}_{\text{hydr}}$ (kcal·mol⁻¹), mean value of changes in hydration free energy when the halogen ions transfer from solvent to halogen-ionic bridges. ⁱ $\Delta\Delta H^{\circ}_{\text{hydr}}$ (kcal·mol⁻¹), mean value of changes in hydration enthalpy when the halogen ions transfer from solvent to halogen-ionic bridges. ^j $-T\Delta\Delta S^{\circ}_{\text{hydr}}$ (kcal·mol⁻¹), mean value of changes in hydration entropy when the halogen ions transfer from solvent to halogen-ionic bridges.

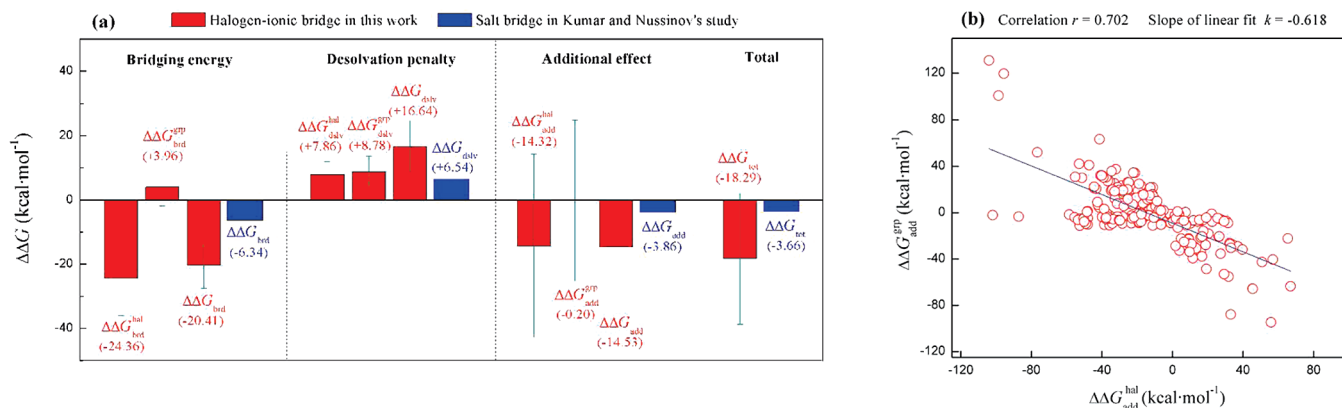


Figure 9. (a) Average energy terms in 241 halogen-ionic bridges formed in monomeric proteins and their comparisons with that in 222 salt bridges. The error bar ranges within \pm standard deviation. (b) Negative linear correlation with a correlation coefficient $r = 0.702$ between $\Delta\Delta G_{\text{add}}^{\text{hal}}$ and $\Delta\Delta G_{\text{add}}^{\text{grp}}$.

continuum electrostatic analysis. Only monomeric proteins are considered here because our primary interest is the effect of halogen-ionic bridges on protein folding rather than on protein binding. The detailed information about these selected halogen-ionic bridges and their electrostatic energy terms are collected in Supporting Information, Table S5.

A total of 204 (84.6%) out of the 241 halogen-ionic bridges in our data set are stabilizing ($\Delta\Delta G_{\text{tot}} < 0$). This stabilization energy $\Delta\Delta G_{\text{tot}}$, as shown in Figure 9a, is a compromise between the favorable electrostatic interaction within the bridge ($\Delta\Delta G_{\text{brd}} < 0$) as well as the interaction of the bridge with the charges in the rest of the protein ($\Delta\Delta G_{\text{add}} < 0$) and the unfavorable desolvation penalty ($\Delta\Delta G_{\text{dslv}} > 0$) of the polar/charged bridge due to its burial in low-dielectric protein interior. On average, halogen-ionic bridge formations incur a desolvation penalty $\Delta\Delta G_{\text{dslv}}$ of $+16.54$ kcal·mol⁻¹. This penalty is over paid by the bridging energy term $\Delta\Delta G_{\text{brd}}$ of -20.41 kcal·mol⁻¹. The electrostatic interactions of halogen-ionic bridges with the rest of the proteins are generally attractive with an average $\Delta\Delta G_{\text{add}}$ of -14.53 kcal·mol⁻¹, which assists the bridging energy term to ultimately overcome the desolvation energy penalty, making the halogen-ionic bridge stabilizing. This free energy profile presented here for halogen-ionic bridge is coincident with that proposed previously for salt bridge,⁶³ but each energy term in the halogen-ionic bridge is much greater than that in the salt bridge (Figure 9a). Consequently, the stabilization effect of halogen-ionic bridge appears to be quite significant in comparison to that of salt bridge, given by the substantial difference in their total stabilization energies $\Delta\Delta G_{\text{tot}}$ (-18.29 for halogen-ionic bridge vs -3.66 kcal·mol⁻¹ for salt bridge).

Breaking down of these energy terms in halogen-ionic bridge into two parts separately associated with the halogen ion and the bridging protein groups could provide a further insight into the energy composition of the bridge. From Figure 9a, it is seen that: (i) the bridging energy, $\Delta\Delta G_{\text{brd}}$, is made up of a dominant attraction term of halogen ion with its interacting groups ($\Delta\Delta G_{\text{brd}}^{\text{hal}} = -24.36$ kcal·mol⁻¹) and a marginal repulsion term among these groups ($\Delta\Delta G_{\text{brd}}^{\text{grp}} = +3.96$ kcal·mol⁻¹); (ii) the desolvation penalty, $\Delta\Delta G_{\text{dslv}}$, is the sum of two nearly equivalent terms as $+7.86$ kcal·mol⁻¹

accounting for the desolvation of halogen ion ($\Delta\Delta G_{\text{dslv}}^{\text{hal}}$) and $+8.78$ kcal·mol⁻¹ for desolvating protein groups ($\Delta\Delta G_{\text{dslv}}^{\text{grp}}$). As can be seen, the value of desolvation penalty for bridging halogen ions calculated using the continuum electrostatic approach is significantly lesser than those obtained from the empirical additive model, as described earlier (see Table 3). This is because the additive model only gives consideration in the “net” hydration effect of halogen ions. It equals the move of halogen ions from a solvent to a gas-phase condition, regardless of the fact that the dielectric constant in protein interiors is actually greater than 1 and the folded proteins are not of infinite extent, so the halogen ions in bridging state can also interact with solvent;⁶² and (iii) the additional effect, $\Delta\Delta G_{\text{add}}$, seems to have arisen from a strong electrostatic attraction between the halogen ion and the rest of the protein ($\Delta\Delta G_{\text{add}}^{\text{hal}} = -14.32$ kcal·mol⁻¹) and from a quite weak term of the bridging groups interacting with the rest of the protein ($\Delta\Delta G_{\text{add}}^{\text{grp}} = -0.20$ kcal·mol⁻¹). However, one should beware of this statement, considering that these two mean quantities are accompanied with large standard deviations, indicating a significant variance within the sample scatters. In fact, most of $\Delta\Delta G_{\text{add}}^{\text{hal}}$ and $\Delta\Delta G_{\text{add}}^{\text{grp}}$ in our data set fall into a wide scope ranging from -80 to $+80$ kcal·mol⁻¹, with few even getting more than $+120$ kcal·mol⁻¹. Moreover, there exists a negative linear correlation between $\Delta\Delta G_{\text{add}}^{\text{hal}}$ and $\Delta\Delta G_{\text{add}}^{\text{grp}}$ (Figure 9b), well reflecting the oppositely charged feature of halogen ion and its bridging groups in interaction with the same charges in the protein region out of the bridge.

3.7. Real Biomolecular System. IV. QM/MM Calculation. Recently, several specific intermolecular forces involved in ligand recognition and binding by protein receptors have been investigated in detail by means of the hybrid QM/MM methodology.^{82–84} These works manifested that, if reasonably collocated with a MM context, it is possible to apply the expensive QM method to treat the nonbonding interactions of interest in the whole biomacromolecular framework. In order to give quantitative insight into the role and significance of halogen-ionic bridges in protein–ligand recognition and to explore their relevance to rational drug design, we herein addressed an ONIOM-based QM/MM study on two paradigms of chlorine-ionic bridges function-

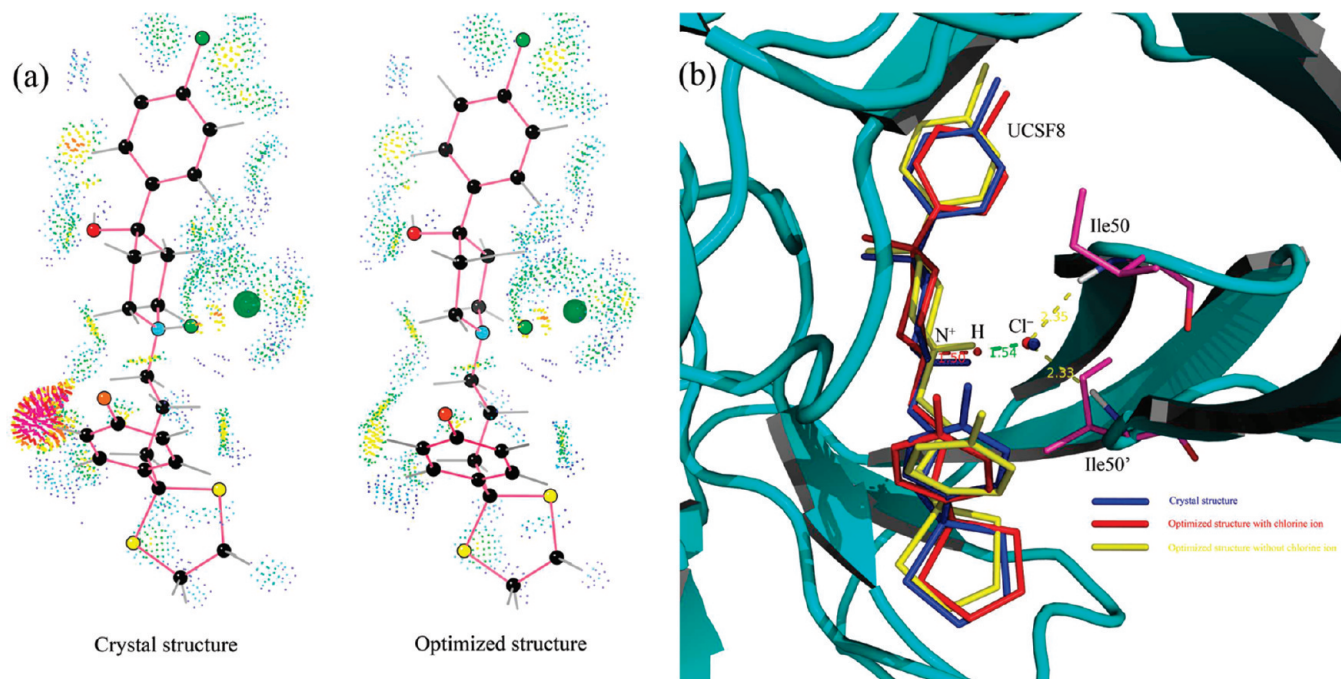


Figure 10. Stereoview of ONIOM-optimized UCSF8 structures in HIV-1 PR Q7K binding pocket. (a) Visualizing vdW clashes between UCSF8 and Q7K in crystal and optimized complex structures. Blue dots for wide contacts (>0.25 Å), green or yellow dots for good contacts (green for close contacts and yellow for slight overlaps, <0.2 Å), and red spikes for bad overlaps (≥0.4 Å). (b) Superposition of the chlorine ion-containing (red) and chlorine ion-removed (yellow) ligands to crystal structure (blue).

alizing to the binding of inhibitors by their targets. The first one is in the earlier mentioned complex of HIV-1 protease (HIV-1 PR) with a nonpeptide inhibitor UCSF8. Here, we used the crystal structure of its mutant, i.e. HIV-1 PR Q7K-UCSF8 complex, which has a similar inhibition profile ($K_i = 15 \mu\text{M}$), similar kinetic parameters but higher resolution level (solved at 1.9 Å) when compared to the wild type, as template to perform QM/MM analysis (PDB: 2aid).²⁶ The second is located at the binding interface of 3-phosphoinositide-dependent kinase 1 (PDK1) with its selective inhibitor BX-320, an aminopyrimidine derivative which can specifically bind to the catalytic domain of PDK1 at a nmol level of affinity ($\text{IC}_{50} = 39 \text{ nM}$) (PDB: 1z5m).⁸⁵ These two complex structures were submitted to an ONIOM minimization procedure, as described in Section 2.1, followed by single-point energy analyses of the optimized model layers using the rigorous MP2/aug-cc-pVDZ theory. For the purpose of comparison, halide anion-removed versions of these two complexes were also analyzed in the same way.

HIV-1 PR Q7K-UCSF8 Complex. Although the crystal structure of this complex was solved at a higher resolution level (1.9 Å), the larger thermal B -factors for atoms of UCSF8 ($\langle B \geq 68.6 \text{ \AA}^2$) versus all atoms in the protein ($\langle B \geq 30.3 \text{ \AA}^2$) suggest that the inhibitor position was not determined quite precisely.²⁶ This point can be validated by detecting van der Waals (vdW) clashes between Q7K and UCSF8 in both the complex crystal structure and the ONIOM-optimized structure. The small probe technique implemented in the PROBE program⁸⁶ was employed to fulfill this purpose, and the result is a graphic diagram visualizing the distribution of collisions around the UCSF8. From Figure 10a, it should be appreciated here that ONIOM

optimization can give a substantial refinement for the active region of the complex crystal structure, as shown by the fact that most of the bad overlaps at the crystal interface were eliminated after the ONIOM minimization procedure. The optimized UCSF8 conformations, with or without chlorine ion, are superposed on the crystal structure (Figure 10b), from which the root-mean-square deviations (RMSDs) of chlorine ion-containing and chlorine ion-removed ligand structures relative to the crystal one were computed to be, respectively, 1.18 and 1.97 Å. The former is far below the X-ray diffraction resolution of the studied crystal, whereas the latter is above of the resolution, indicating that the optimized model structures should be reliable and the absence of chlorine ion would throw a considerable effect on the ligand arrangement in Q7K binding pocket. Noteworthy, in the optimized structure, the N–H bond of UCSF8 amine moiety is elongated remarkably as much as to 1.50 Å, manifesting that there exists a significant trend of proton transfer toward the chlorine ion, which is consistent with that found earlier in small model systems. Energy analysis further revealed a noticeable effect of the proton transfer contributing to the complex stabilization. Interaction energy between the chlorine ion and UCSF8 was predicted to be $-121.10 \text{ kcal}\cdot\text{mol}^{-1}$, this value is far more than that when a water molecule is placed at the same position of the chlorine ion ($-15.97 \text{ kcal}\cdot\text{mol}^{-1}$). In addition, the chlorine ion is shown to be also effective in interaction with the residues Ile50 and Ile50' of Q7K, on account of the strong intermolecular potential of $-40.03 \text{ kcal}\cdot\text{mol}^{-1}$. In conclusion, the chlorine-ionic bridge should be important in assisting the specific binding of Q7K by UCSF8 and in maintaining the complex conformation and stability.

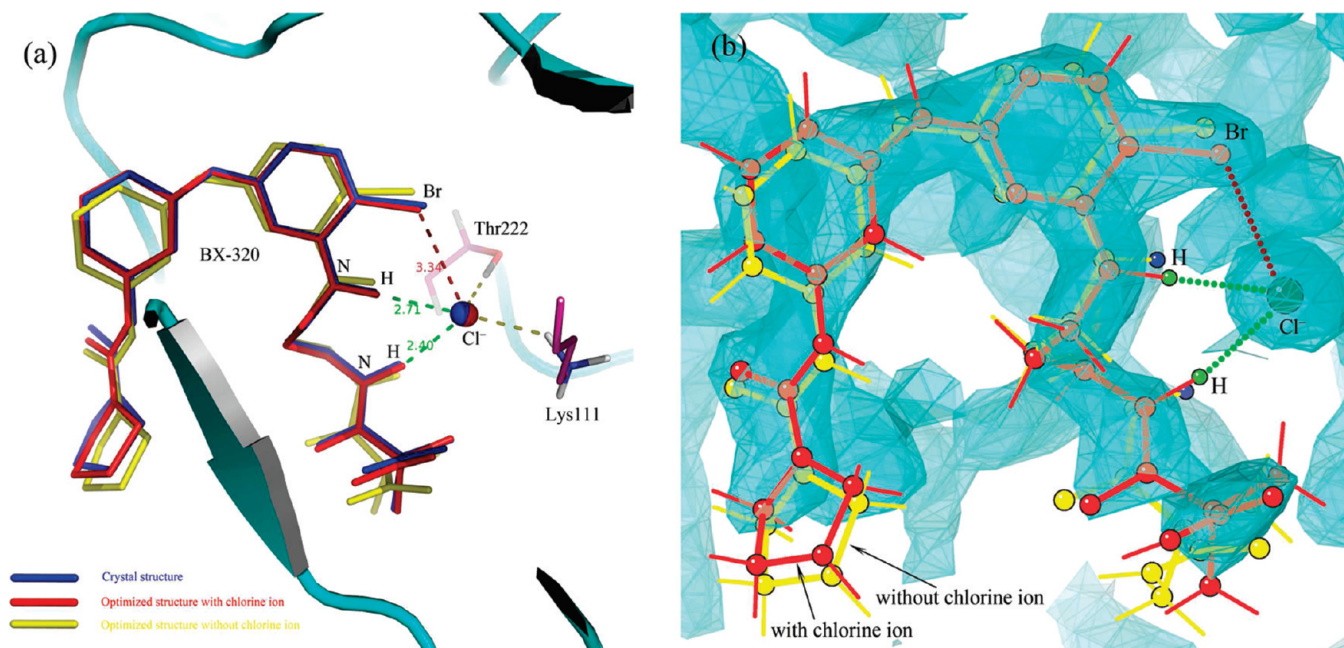


Figure 11. Stereoview of ONIOM-optimized BX-320 structures in PDK1 binding pocket. (a) Superposition of the chlorine ion-containing (red) and chlorine ion-removed (yellow) structures to crystal structure (blue). (b) Superposition of chlorine ion-containing (red) with chlorine ion-removed (yellow) structures in 2Fo–Fc electron-density map (shown contoured at 2.2σ).

PDK1-BX-320 Complex. ONIOM optimization of the huge BX-320 molecule in the PDK1 binding pocket was an exhaustive process, but the resulting BX-320 structure had only a little conformational change relative to original crystal structure (RMSD = 0.73 Å). As shown in Figure 11a, the optimized BX-320 structure is nearly perfectly superposed on the crystal, with slight fluctuations over whole molecular heavy atoms. Each of the two N–H bonds in the flexible chain of BX-320 forms a typical ionic hydrogen bond with the chlorine ion that is bound to the residues Lys111 and Thr222 of PDK1. Intriguingly, the bromine atom on the BX-320 pyrimidine ring seems to be in weak interaction with the chlorine ion through a nonlinear halogen bond,^{87–89} as claimed by their interatomic distance being shorter than the sum of respective van der Waals radii ($3.34 < 3.66$ Å)^{57,90} and their angle $\angle(\text{C}-\text{Br}\cdots\text{Cl}^-)$ meeting the criterion defined by Auffinger et al. for biological halogen bonds ($121.3^\circ > 120^\circ$).⁷² Topological analysis of the electron density in the optimized model layer confirmed the existence of a bond path linking the nuclei of Br and Cl⁻. However, the value (0.0132 au) of electron density ρ_b at the BCP for this interaction was predicted to be much smaller than that for two N–H \cdots Cl⁻ interactions (0.0365 and 0.0313 au), implying a very weak halogen bond in comparison to the strong ionic hydrogen bonds in this model. The Br \cdots Cl⁻ halogen bond can be, therefore, considered as secondary interaction contributions to the H \cdots Cl⁻ hydrogen bonds that conduct the formation of the chlorine-ionic bridge. At the MP2/aug-cc-pVDZ level, the interaction energies of the chlorine ion with BX-320 and PDK1 were calculated to be -34.02 and -131.62 kcal \cdot mol⁻¹, respectively. This noticeable energy level involved in the chlorine-ionic bridge should contribute considerably to the binding affinity of BX-320, though the desolvation penalty is not deducted from the QM-calculated interaction energies. The functionality of the chlorine ion in

the PDK1 active site can be intuitively characterized by comparing the optimized conformations of chlorine ion-containing and chlorine ion-removed complexes. It can be seen from Figure 11b that BX-320 structure exhibits an obvious motion when the chlorine ion is taken off from the complex, leading to a large RMSD (1.94 Å) relative to that optimized with the chlorine ion. In addition, the chlorine ion-containing structure can be fitted into the 2Fo–Fc electron-density map fairly well, but the chlorine ion-removed one, especially at its two ends, departs from the 2.2σ contour appreciably, indicating a nonignorable effect of the chlorine ion on the native architecture of this system.

4. Conclusions

The main aim of this study is to prove the existence and significance of the putative halogen-ionic bridges in biomolecular systems. To achieve this, we present a comprehensive investigation on the geometrical profile and energy landscape of biological interactions involving halide anions. High-level ab initio calculations on small model systems preliminarily unveil the noticeable stabilization and the typical bonding character of halogen ion complexes with polar and charged groups. Database surveys of massive crystal structures deposited in the PDB further reveal a considerable number of geometrically preferential contacts between the nonbonded halogen ions and the electrophilic moieties of proteins, nucleic acids, and small ligands; these contacts are used to define a distinct data set consisting of 6406 biological halogen-ionic bridges. Continuum electrostatic analyses and hybrid quantum mechanics/molecular mechanics (QM/MM) examinations ultimately give a quantitative pronouncement for the important role of halogen-ionic bridges in conferring stability and specificity for protein folding and protein–ligand binding. All of these forebode that the halogen-ionic bridges should widely exist in

and solidly functionalize to biomolecules. We, therefore, expect that this newly proposed halide motif, with respect to its substantial magnitude and marked stabilization in biological context, could be exploited as a novel and versatile tool for rational drug design and bioengineering.

Acknowledgment. This work was supported by the State Key Laboratory of Trauma, Burns, and Combined Injury Foundation (no. SKLKF200904).

Supporting Information Available: A classification list of the 6406 biological halogen-ionic bridges retrieved from the current PDB database (January, 2010 release) and the calculated electrostatic energy terms for the 241 high-quality halogen-ionic bridges are tabulated. The six protein moieties used in continuum electrostatic calculations and their PARSE parameters as well as comparison of intermolecular potentials to ideal and real Coulombic energies for some small model complexes are shown. Tables S1–S5; Figures S1–S3. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Hofmeister, F. *Arch. Exp. Pathol. Pharmacol.* **1888**, *24*, 247–260.
- Marcus, Y. *Chem. Rev.* **2009**, *109*, 1346–1370.
- Zhang, Y.; Cremer, P. S. *Curr. Opin. Chem. Biol.* **2006**, *10*, 658–663.
- Takashima, K.; Riveros, J. M. *Mass Spectrom. Rev.* **1998**, *17*, 409–430, and references therein.
- Kovács, A.; Varga, Z. *Coord. Chem. Rev.* **2006**, *250*, 710–727, and references therein.
- Arshadi, M.; Yamdagni, R.; Kebarle, P. *J. Phys. Chem.* **1970**, *74*, 1475–1482.
- Caldwell, G.; Kebarle, P. *J. Am. Chem. Soc.* **1984**, *106*, 967–969.
- Hiraoka, K.; Mizuse, S.; Yamabe, S. *J. Phys. Chem.* **1988**, *92*, 3943–3957.
- Bogdanov, B.; Peschke, M.; Tonner, D. S.; Szulejko, J. E.; McMahon, T. B. *Int. J. Mass Spectrom.* **1999**, *187*, 707–725.
- Klopper, W.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Duijneveldt, F. B. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2227–2234.
- Xantheas, S. S.; Dunning, T. H., Jr. *J. Phys. Chem.* **1994**, *98*, 13489–13497.
- Johnson, M. S.; Kuwata, K. T.; Wong, C. K.; Okumura, M. *Chem. Phys. Lett.* **1996**, *260*, 551–557.
- Choi, J. H.; Kuwata, K. T.; Cao, Y. B.; Okumura, M. *J. Phys. Chem. A* **1998**, *102*, 503–507.
- Kim, J.; Lee, H. M.; Suh, S. B.; Majumdar, D.; Kim, K. S. *J. Chem. Phys.* **2000**, *113*, 5259–5272.
- Bogdanov, B.; McMahon, T. B. *J. Phys. Chem. A* **2000**, *104*, 7871–7880.
- Vinogradov, S. N. *Int. J. Pept. Protein Res.* **1979**, *14*, 281–289.
- Zhang, W.; Piculell, L.; Nilsson, S. *Macromolecules* **1992**, *25*, 6165–6172.
- Washabaugh, M. W.; Collins, K. D. *J. Biol. Chem.* **1986**, *261*, 12477–12485.
- Chen, X.; Yang, T.; Kataoka, S.; Cremer, P. S. *J. Am. Chem. Soc.* **2007**, *129*, 12272–12279.
- Lund, M.; Vrbka, L.; Jungwirth, P. *J. Am. Chem. Soc.* **2008**, *130*, 11582–11583.
- (a) Ninham, B. W.; Yaminsky, V. *Langmuir* **1997**, *13*, 2097–2108. (b) Böstrom, M.; Tavares, F. W.; Bratko, D.; Ninham, B. W. *J. Phys. Chem. B* **2005**, *109*, 24489–24494. (c) Boström, M.; Lonetti, B.; Fratini, E.; Baglioni, P.; Ninham, B. W. *J. Phys. Chem. B* **2006**, *110*, 7563–7566.
- (a) Lund, M.; Jungwirth, P.; Woodward, C. E. *Phys. Rev. Lett.* **2008**, *100*, 258105. (b) Lund, M.; Vácha, R.; Jungwirth, P. *Langmuir* **2008**, *24*, 3387–3391.
- Heyda, J.; Hrobárik, T.; Jungwirth, P. *J. Phys. Chem. A* **2009**, *113*, 1969–1975.
- Meyer, E. *Protein Sci.* **1992**, *1*, 1543–1562.
- Sun, Y.; Kollman, P. *J. Phys. Chem.* **1996**, *100*, 6760–6763.
- Rutenber, E.; Fauman, E. B.; Keenan, R. J.; Fong, S.; Furth, P. S.; Ortiz de Montellano, P. R.; Meng, E.; Kuntz, I. D.; DeCampn, D. L.; Saltoll, R.; Rosb, J. R.; Craik, C. S.; Stroud, R. M. *J. Biol. Chem.* **1993**, *268*, 15343–15346.
- Suresh, C. H.; Vargheese, A. M.; Vijayalakshmi, K. P.; Mohan, N.; Koga, N. *J. Comput. Chem.* **2008**, *29*, 1840–1849.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
- Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford University Press: Oxford, U.K., 1990.
- Reed, A. E.; Curitts, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 889–926.
- Chesnut, D. B.; Moseley, R. W. *Theor. Chim. Acta.* **1969**, *13*, 230–248.
- Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.
- Glukhovtsev, M. N.; Pross, A.; Radom, L. *J. Am. Chem. Soc.* **1995**, *117*, 2024–2032.
- Sanov, A.; Faeder, J.; Parson, R.; Lineberger, W. C. *Chem. Phys. Lett.* **1999**, *313*, 812–819.
- Cioslowski, J.; Piskorz, P. *Chem. Phys. Lett.* **1996**, *255*, 315–319.
- Asaduzzaman, A. M.; Schreckenbach, G. *Theor. Chem. Acc.* **2009**, *122*, 119–125.
- Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.
- Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- Comell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620–9631.
- Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

- (43) Zhou, P.; Lv, J.; Zou, J.; Tian, F.; Shang, Z. *J. Struct. Biol.* **2010**, *169*, 172–182.
- (44) Zhou, P.; Zou, J.; Tian, F.; Shang, Z. *J. Chem. Inf. Model.* **2009**, *49*, 2344–2355.
- (45) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2003.
- (46) Biegler-Konig, F.; Schoenbohm, J. *AIM2000*, 2nd ed.; Buro fur Innovative Software: Bielefeld, Germany, 2002.
- (47) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. *NBO 5.0*; Theoretical Chemistry Institute, University of Wisconsin: Madison, WI, 2001.
- (48) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Jr *Proteins* **2009**, *77*, 778–795.
- (49) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.
- (50) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol.* **1999**, *285*, 1735–1747.
- (51) Zhou, P.; Tian, F.; Lv, F.; Shang, Z. *Proteins* **2009**, *76*, 151–163.
- (52) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins* **2008**, *73*, 765–783.
- (53) Zhao, Y.; Cheng, T.; Wang, R. *J. Chem. Inf. Model.* **2007**, *47*, 1379–1385.
- (54) Sanner, M. F.; Olson, A. J.; Spehner, J.-C. *Biopolymers* **1996**, *38*, 305–320.
- (55) Rother, K.; Hildebrand, P. W.; Goede, A.; Gruening, B.; Preissner, R. *Nucleic Acids Res.* **2009**, *37*, D393–D395.
- (56) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. *J. Mol. Biol.* **1999**, *290*, 253–266.
- (57) Shannon, R. D. *Acta Crystallogr.* **1976**, *A32*, 751–767.
- (58) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086–3090.
- (59) Klots, C. E. *J. Phys. Chem.* **1981**, *85*, 3585–3588.
- (60) Rocchia, W.; Alexov, E.; Honig, B. *J. Phys. Chem.* **2001**, *105*, 6507–6514.
- (61) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.
- (62) Hendsch, Z. S.; Tidor, B. *Protein Sci.* **1994**, *3*, 211–226.
- (63) Kumar, S.; Nussinov, R. *J. Mol. Biol.* **1999**, *293*, 1241–1255.
- (64) Naray-Szabo, G.; Ferenczy, G. G. *Chem. Rev.* **1995**, *95*, 829–847.
- (65) Meot-Ner, M. *Chem. Rev.* **2005**, *105*, 213–284.
- (66) Nakanishi, W.; Hayashi, S.; Narahara, K. *J. Phys. Chem. A* **2008**, *112*, 13593–13599.
- (67) Wiberg, K. B. *Tetrahedron* **1968**, *24*, 1083–1096.
- (68) Reed, A. E.; Schleyer, P. *Inorg. Chem.* **1988**, *27*, 3969–3987.
- (69) Chesnut, D. B. *J. Chem. Theory Comput.* **2008**, *4*, 1637–1642.
- (70) Prabakaran, P.; Singarayan, M. G. *Chem. Lett.* **2004**, *33*, 1640–1641.
- (71) Rajagopal, S.; Vishveshwara, S. *FEBS J.* **2005**, *272*, 1819–1832.
- (72) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 16789–16794.
- (73) Baldwin, R. L.; Rose, G. D. *Trends Biochem. Sci.* **1999**, *24*, 77–84.
- (74) Tsai, C. J.; Xu, D.; Nussinov, R. *Folding Des.* **1998**, *3*, R71–R80.
- (75) Osawa, S. *Annu. Rev. Biochem.* **1968**, *37*, 109–130.
- (76) Lu, Y.; Wang, R.; Yang, C. Y.; Wang, S. *J. Chem. Inf. Model.* **2007**, *47*, 668–675.
- (77) Srinivasan, R.; Rose, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14258–14263.
- (78) Sun, D. P.; Sauer, U.; Nicholson, H.; Matthews, B. W. *Biochemistry* **1991**, *30*, 7142–7153.
- (79) Dao-pin, S.; Anderson, D. E.; Baase, W. A.; Dahlquist, F. W.; Matthews, B. W. *Biochemistry* **1991**, *30*, 11521–11529.
- (80) Waldburger, C. D.; Schildbach, J. F.; Sauer, R. T. *Nat. Struct. Biol.* **1995**, *2*, 122–128.
- (81) Noguchi, T.; Matsuda, H.; Akiyama, Y. *Nucleic Acids Res.* **2001**, *29*, 219–220.
- (82) Lu, Y.; Shi, T.; Wang, Y.; Yang, H.; Yan, X.; Luo, X.; Jiang, H.; Zhu, W. *J. Med. Chem.* **2009**, *52*, 2854–2862.
- (83) Lu, Y.; Wang, Y.; Xu, Z.; Yan, X.; Luo, X.; Jiang, H.; Zhu, W. *J. Phys. Chem. B* **2009**, *113*, 12615–12621.
- (84) Alzate-Morales, J. H.; Caballero, J.; Jague, A. V.; Nilo, F. D. G. *J. Chem. Inf. Model.* **2009**, *49*, 886–899.
- (85) Feldman, R. I.; Wu, J. M.; Polokoff, M. A.; Kochanny, M. J.; Dinter, H.; Zhu, D.; Biroc, S. L.; Alicke, B.; Bryant, J.; Yuan, S.; Buckman, B. O.; Lentz, D.; Ferrer, M.; Whitlow, M.; Adler, M.; Finster, S.; Chang, Z.; Arnaiz, D. O. *J. Biol. Chem.* **2005**, *280*, 19867–19874.
- (86) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol.* **1999**, *285*, 1711–1733.
- (87) Politzer, P.; Lane, P.; Concha, M. C.; Ma, Y.; Murray, J. S. *J. Mol. Model.* **2007**, *13*, 305–311.
- (88) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. *J. Mol. Model.* **2007**, *13*, 291–296.
- (89) Murray, J. S.; Lane, P.; Politzer, P. *J. Mol. Model.* **2009**, *15*, 723–729.
- (90) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.

A Non-Orthogonal Block-Localized Effective Hamiltonian Approach for Chemical and Enzymatic Reactions

Alessandro Cembran, Apirak Payaka, Yen-lin Lin, Wangshen Xie, Yirong Mo,^{*,†}
Lingchun Song,^{*} and Jiali Gao^{*}

*Department of Chemistry, Digital Technology Center and Supercomputing Institute,
University of Minnesota, Minneapolis, Minnesota 55455, and Department of
Chemistry, Western Michigan University, Kalamazoo, Michigan 49008*

Received March 28, 2010

Abstract: The effective Hamiltonian–molecular orbital and valence bond (EH-MOVB) method based on nonorthogonal block-localized fragment orbitals has been implemented in the program CHARMM for molecular dynamics simulations of chemical and enzymatic reactions, making use of semiempirical quantum mechanical models. Building upon *ab initio* MOVB theory, we make use of two parameters in the EH-MOVB method to fit the barrier height and the relative energy between the reactant and product state for a given chemical reaction to be in agreement with experimental or high-level *ab initio* or density functional results. Consequently, the EH-MOVB method provides a highly accurate and computationally efficient QM/MM model for dynamics simulation of chemical reactions in solution. The EH-MOVB method is illustrated by examination of the potential energy surface of the hydride transfer reaction from trimethylamine to a flavin cofactor model in the gas phase. In the present study, we employed the semiempirical AM1 model, which yields a reaction barrier that is more than 5 kcal/mol too high. We use a parameter calibration procedure for the EH-MOVB method similar to that employed to adjust the results of semiempirical and empirical models. Thus, the relative energy of these two diabatic states can be shifted to reproduce the experimental energy of the reaction, and the barrier height is optimized to reproduce the desired (accurate) value by adding a constant to the off-diagonal matrix element. The present EH-MOVB method offers a viable approach to characterizing solvent and protein-reorganization effects in the realm of combined QM/MM simulations.

1. Introduction

Combined quantum mechanical and molecular mechanical (QM/MM) methods offer an excellent opportunity for studying chemical and electron transfer reactions in solution and in biological systems.^{1–3} In principle, the accuracy of combined QM/MM potentials can be systematically improved; however, it is still time-demanding to carry out QM/MM simulations using *ab initio* wave function theory (WFT) or density functional theory (DFT) for subsystems consisting of more than 100 atoms in the QM region. Consequently, it

is useful to develop efficient QM/MM techniques that can be made accurate for specific chemical and biomolecular applications, yet sufficiently fast for extensive conformational sampling. Aside from the brute force approach by increasing the level of theory and the size of basis set, there are two other ways to achieve this goal. The first is to parametrize purely empirical energy functions to model a specific process,^{4,5} and the second is to parametrize quantum mechanical models against experimental data with specific reaction parameters (SRP) for a given class of reactions.^{6–9} In this article, we describe an effective Hamiltonian approach based on the molecular orbital-valence bond (MOVB) theory developed in our laboratories for the treatment of reactive potential surfaces of reactions.^{10–12} In particular, we illustrate

^{*} To whom correspondence should be addressed. E-mail: yirong.mo@wmich.edu (Y.M.), lcsong2007@gmail.com (L.S.), gao@jialigao.org (J.G.).

[†] Western Michigan University.

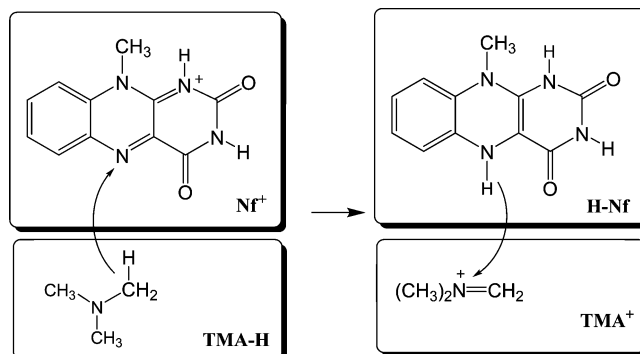
this novel QM/MM technique in the hydride transfer reaction from trimethyl ammonium ion to flavin cofactor.

The mixed molecular orbital and valence bond (MOVB) theory,^{10,11} initially developed at the *ab initio* level and recently extended to multistate density functional theory (called MSDFT or, equivalently, VBDF),¹³ is designed to treat reactive potential energy surfaces for chemical reactions and electron transfer processes. In this approach, the whole process is described with two or more resonating configurations, including the reactant and product states. In each specific state, molecular orbitals (MOs) are strictly localized within individual fragments of a molecular system.^{14–20} However, the block-localized molecular orbitals (BLMOs) are still delocalized within each orbital block, making the MOVB method extremely efficient in comparison with *ab initio* valence bond theory.^{21–25} Key features of the MOVB and MSDFT theories include (1) that the BLMOs (or block-localized Kohn–Sham orbitals)¹³ within each fragment are orthogonal, which makes it computationally efficient, and (2) that the BLMOs between different fragments are nonorthogonal,^{11,13,20} which retains important characters of valence bond (VB) theory. In the limiting case in which there is one fragment, MOVB and MSDFT reduce exactly to the standard Hartree–Fock theory and Kohn–Sham DFT, respectively.

Recently, we introduced an effective Hamiltonian MOVB approach,²⁶ in which the *ab initio* electronic matrix elements are adjusted to yield accurate barrier height and reaction enthalpy. This approach has an apparent similarity in the “calibration” process used to adjust the barrier height and the energy of reaction in semiempirical or empirical valence bond models,^{27–32} although the theory and algorithm of MOVB are based on *ab initio* WFT and DFT approaches to define VB electronic configuration states. Effective Hamiltonian approaches are widely used in many different areas.^{33–40} A major advantage of the EH-MOVB approach is that all VB matrix elements, including off-diagonal terms, are determined by an electronic structure method, which depends explicitly on all degrees of freedom in the system. In the empirical and semiempirical valence bond approaches, typically a simple function, depending on one degree of freedom, or a constant is used to treat the off-diagonal matrix elements in a VB-like Hamiltonian.^{27,30,31,41} Note that empirical multiconfigurational models have been described to fit the energy, gradient, and Hessian of *ab initio* potential surfaces^{40,42–44} using Gaussian and polynomial functions^{45,46} or Shepard interpolation.^{47–49}

In this paper, we show that the EH-MOVB method can be constructed using semiempirical QM models such as the Austin model 1 (AM1),⁵⁰ parametrization model 3 (PM3),⁵¹ or Recife model 1 (RM1)⁵² to yield the barrier height for a chemical reaction in agreement with experiments or with *ab initio* results. In the following, we first present the theoretical background, followed by computational details. Results and discussions are presented next. Finally, the paper concludes with a summary of the major findings of this study and future perspectives.

Scheme 1. Schematic Representation of the Block-Localization of Molecular Orbitals within Individual Molecular Fragments for the Reactant Diabatic State (left) and the Product Diabatic State (right) for the Hydride Transfer Reaction between Trimethylamine (TMA-H) and a Model for the Flavin Cofactor (Nf^+)^a



^a Atoms and charges in each rectangle specify the molecular block defined by the corresponding Lewis structure within which molecular orbitals are localized. The antisymmetric wave function constructed from the two blocks on the left-hand side of the arrow, **TMA-H** and **Nf⁺**, defines the reactant diabatic state, whereas that for the blocks on the right-hand side, **TMA⁺** and **H-Nf**, define the product diabatic state.

2. Method

A. The Mixed Molecular Orbital and Valence Bond (MOVB) Theory. In MOVB,^{10–12,21} we use one Slater determinant wave function constructed using nonorthogonal block-localized molecular orbitals (BLMO) to define the reactant and product configurations. These electronic configurations are called diabatic states. The use of localized orbitals within molecular fragments has been explored by many groups in different applications such as reducing basis set superposition errors in weakly bound complexes,^{17–19,53,54} and it has been used in other contexts.^{14–16,55–61} For the hydride transfer reaction between trimethylamine, $(\text{CH}_3)_3\text{N}$ (**TMA-H**), and a flavin cofactor (Nf^+) model (hereafter simply called flavin), the wave function of the reactant diabatic state, $\Psi_r(\mathbf{R})$ (see Scheme 1), is defined by a single Slater determinant wave function in which molecular orbitals are block-localized on the two subsystems:

$$\Psi_r(\mathbf{R}) = \hat{A}\{\chi_r^{\text{TMA-H}} \chi_r^{\text{Nf}^+}\} \quad (1)$$

where \mathbf{R} specifies all Cartesian atomic coordinates of the system and \hat{A} is an antisymmetrization operator. The notations $\chi_r^{\text{TMA-H}}$ and $\chi_r^{\text{Nf}^+}$ in eq 1 specify the products of occupied BLMOs that are defined as linear combinations of atomic orbitals located on atoms in fragments **TMA-H** and **Nf⁺**, respectively (Scheme 1). Similarly, the wave function of the product state (Scheme 1), $\Psi_p(\mathbf{R})$, is expressed as

$$\Psi_p(\mathbf{R}) = \hat{A}\{\chi_p^{\text{TMA}^+} \chi_p^{\text{H-Nf}}\} \quad (2)$$

where $\chi_p^{\text{TMA}^+}$ and $\chi_p^{\text{H-Nf}}$ denote the products of occupied BLMOs expanded over basis orbitals on atoms in fragments **TMA⁺** and **H-Nf**, respectively (Scheme 1).

It is important to note that the MOs within each fragment for each state are constrained to be orthogonal, but they are

nonorthogonal between different fragments.¹¹ Consequently, the MOVb model retains key characteristic features of valence bond theory in the use of nonorthogonal orbitals. The structure of the transformation matrix for the reactant and product states is block-diagonal.

$$\mathbf{C}_r = \begin{pmatrix} \mathbf{C}_r^{\text{TMA-H}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_r^{\text{Nf}^+} \end{pmatrix} \text{ and } \mathbf{C}_p = \begin{pmatrix} \mathbf{C}_p^{\text{TMA}^+} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_p^{\text{H-Nf}} \end{pmatrix} \quad (3)$$

where \mathbf{C}_r and \mathbf{C}_p are the matrices of molecular orbital coefficients for the reactant and product states, respectively.¹¹ Note that the dimensions of the matrix elements in eq 3 are different as the hydride atom is grouped in different blocks. The total number of electrons within each fragment of each diabatic state is also fixed according to the corresponding Lewis structure (Scheme 1), and there is no chemical bond between the two fragments in each state.

In the present EH-MOVb model employing a semiempirical method, a special situation must be considered because of the neglect of diatomic differential overlap (NDDO) approximation.⁶² The two-center, one-electron integral between two atoms that are located in different blocks is not included in the Fock matrix of either block, and it must be treated specifically. Note that these integrals are formally ignored in the NDDO approximation, but they are also treated as an exception in standard semiempirical methods because of the importance of these integrals in describing chemical bonding.⁵⁰ In MOVb, when the two bonding atoms involved in bond formation are in different molecular blocks (fragments), the two-center, one-electron integral is not treated by the standard semiempirical model, and the exclusion of this type of resonance integral affects the bonding properties as the chemical bonds are partially formed and broken across different blocks (fragments) at the transition state. Consequently, we need to include these resonance integrals for the corresponding bonds in the diabatic energy term as follows. For the reactant state in the present hydride transfer reaction, the reactant state diabatic energy is given as follows:

$$H_{rr}(\mathbf{R}) = \langle \Psi_r(\mathbf{R}) | H | \Psi_r(\mathbf{R}) \rangle + \alpha_{\text{HN}} S_{\text{HN}} \frac{1}{2} [\beta_s(\text{H}) + \beta_{\text{sp}}(\text{N})] \quad (4)$$

and the energy for the product state is

$$H_{pp}(\mathbf{R}) = \langle \Psi_p(\mathbf{R}) | H | \Psi_p(\mathbf{R}) \rangle + \alpha_{\text{CH}} S_{\text{CH}} \frac{1}{2} [\beta_s(\text{H}) + \beta_{\text{sp}}(\text{C})] \quad (5)$$

where S_{HN} and S_{CH} are the s-type overlap integrals between the acceptor nitrogen atom and the transferring hydrogen atom, and between the donor carbon atom and the migrating hydrogen atom, specified by the subscripts respectively, and $\beta_{\text{sp}}(\text{X}) = [\beta_s(\text{X}) + 3\beta_p(\text{X})]/4$ at $\text{X} = \text{N}$ or C and the β 's being the standard semiempirical parameters for these atoms.^{50,51} The use of s-type overlap integrals in eqs 4 and 5 is to preserve rotation invariance. In eqs 4 and 5, we treat α_{HN} and α_{CH} as semiempirical parameters, adjusted to yield the corresponding bond distances in agreement with DFT energies at the transition state. These two parameters associated with bonding interactions may also be considered

as EH-MOVb parameters, in addition to the two parameters to adjust diabatic coupling results.

The MOVb wave function for the reactive system is written as a linear combination of the diabatic states.

$$\Phi_g(\mathbf{R}) = a_r \Psi_r(\mathbf{R}) + a_p \Psi_p(\mathbf{R}) \quad (6)$$

where a_r and a_p are the configurational coefficients for the reactant and product diabatic states, respectively.^{15,16,20,26} The potential energy of the adiabatic ground state, $V_g(\mathbf{R})$, is the lower energy root of the secular equation.

$$\begin{vmatrix} H_{rr}(\mathbf{R}) - V(\mathbf{R}) & H_{rp}(\mathbf{R}) - S_{rp}(\mathbf{R})V(\mathbf{R}) \\ H_{pr}(\mathbf{R}) - S_{pr}(\mathbf{R})V(\mathbf{R}) & H_{pp}(\mathbf{R}) - V(\mathbf{R}) \end{vmatrix} = 0 \quad (7)$$

where $V(\mathbf{R})$ is the adiabatic potential energy, $H_{rr}(\mathbf{R})$ and $H_{pp}(\mathbf{R})$ are the Hamiltonian matrix elements for the reactant and product diabatic states, respectively, $H_{rp}(\mathbf{R}) = H_{pr}(\mathbf{R})$ is the exchange integral (off-diagonal matrix element), and $S_{rp}(\mathbf{R}) = S_{pr}(\mathbf{R})$ is the overlap integral between the two diabatic states.

The Hamiltonian matrix elements in eq 7 are given as follows:^{11,13}

$$H_{ab} = S_{ab} \left\{ Tr[(\mathbf{D}_{ab})^T \mathbf{h}] + \frac{1}{2} Tr[(\mathbf{D}_{ab})^T \mathbf{J} \mathbf{D}_{ab}] - \frac{1}{4} Tr[(\mathbf{D}_{ab})^T \mathbf{K} \mathbf{D}_{ab}] + E_{\text{nuc}} \right\} \quad (8)$$

where the subscripts a and b specify either the reactant (r) or the product (p) state or both; E_{nuc} is the nuclear Coulomb energy; S_{ab} and \mathbf{D}_{ab} are the overlap integral and density matrix over nonorthogonal determinant wave functions; and \mathbf{h} , \mathbf{J} , and \mathbf{K} are the standard one-electron, Coulomb, and exchange matrices. It is important to note that eq 8 is a general formula that is valid for *ab initio* and semiempirical WFT as well as for standard Kohn–Sham DFT.¹³ In the latter case, the exchange integral \mathbf{K} is replaced by the exchange-correlation potential.¹³

In reference,²⁰ we described two special situations to optimize the wave function of eq 6. In the first case, which is called the consistent diabatic configurations (CDC) MOVb, both the orbital coefficients (eq 3) and configurational coefficients are optimized as in the multiconfiguration self-consistent field method. An alternative approach is to variationally optimize the reactant and product state separately, followed by optimizing the configurational coefficient in eq 6 with the orbital coefficients kept fixed. The latter configuration interaction procedure is called the variational diabatic configuration (VDC) MOVb to emphasize that the diabatic states are individually optimized. Both CDC and VDC states are useful in condensed phase simulations, although their applications will be addressed in future publications.

B. Effective Hamiltonian MOVb. We aim to develop an efficient (e.g., capable of carrying out nanosecond to microsecond dynamics simulations using the current computer architecture) and accurate (within 1 to 2 kcal/mol of experimental barrier height) QM/MM method for simulation of enzymatic reactions and chemical processes in solution using MOVb. Although *ab initio* MOVb and multistate

VBDFT provide a natural choice, and the former has indeed been applied to a number of condensed phase reactions,^{10–12,59,63} it is still very time-demanding to carry out routine free energy simulations, in which a large number of atoms are treated quantum-mechanically. To this end, we have implemented the MOVb method into the CHARMM package,⁶⁴ based on the NDDO approximations.^{62,65} The present implementation represents a significant advance in combined QM/MM methodology because (a) semiempirical methods are computationally efficient, allowing for statistical mechanical sampling in molecular dynamics simulations, and (b) the computational accuracy can be conveniently achieved using the nonorthogonal block-localized orbital approach described here.²⁶

Experience shows that the qualitative features of the potential surface for chemical processes can be adequately represented by semiempirical models, such as AM1,⁵⁰ PM3,⁵¹ or the self-consistent charge tight-binding density functional algorithm (SCC-DFTB).⁶⁶ Consequently, we define and describe the reactant and product diabatic states using a semiempirical Hamiltonian. The quantitative errors in the computed barrier height and the energy of reaction inherited in the semiempirical method are eliminated by adjusting the EH-MOVb matrix elements²⁶ in a similar way to that in empirical or semiempirical VB models.^{27,30,31,41} It should be realized that all combined QM/MM methods are semiempirical models in that one has to employ empirical potential functions such as the Lennard-Jones terms to approximate the quantum mechanical exchange repulsion and dispersion interactions between the QM and MM regions. Thus, the adjustment of the EH-MOVb matrix elements is no stranger to combined QM/MM methodologies.

Specifically, we introduce a parameter in the off-diagonal Hamiltonian matrix element H_{rp} , which is optimized in order to reproduce the barrier height for a given chemical reaction:

$$H_{rp}^{\text{EH}} = H_{rp} + \gamma_{rp} \quad (9)$$

In eq 9, H_{rp} is the MOVb off-diagonal matrix element that is determined directly (eq 8) using a given semiempirical model, γ_{rp} is a parameter that affects dominantly the computed barrier height, and H_{rp}^{EH} is the total effective Hamiltonian (EH) resonance (exchange) integral. Another formalism that we have explored is to scale the off-diagonal matrix element as follows:²⁶

$$H_{rp}^{\text{EH}} = \xi_{rp} H_{rp} \quad (10)$$

Both options can be useful, depending on the performance of the semiempirical model and the specific reaction considered, and both are available options in our implementation in CHARMM. In eq 9, the resonance integral is shifted by a constant value, whereas the scaling procedure in eq 10 affects the dependence of the resonance integral on the overlap between the reactant and product diabatic states. For the hydride transfer reaction between trimethylamine and flavin, we found that eq 9 yields the best results, and it is employed in the present study.

The second parameter that we introduce in the EH-MOVb model is the adjustment of the relative energy between the

reactant and product diabatic states. Thus, if necessary, the diagonal MOVb matrix element for the product state, H_{pp} , is shifted by an amount of $\Delta\varepsilon$ to yield the desired energy of reaction for the process of interest:

$$H_{pp}^{\text{EH}} = H_{pp} + \Delta\varepsilon \quad (11)$$

The value of the parameter $\Delta\varepsilon$ is readily estimated as follows:

$$\Delta\varepsilon = \Delta E_{\text{expt}} - \Delta E_{\text{MOVb}} \quad (12)$$

where $\Delta E_{\text{MOVb}} = H_{pp}(\mathbf{R}_p) - H_{rr}(\mathbf{R}_r)$, which is the relative energy of the unshifted reactant and product diabatic state at their corresponding equilibrium geometries \mathbf{R}_r and \mathbf{R}_p , and ΔE_{expt} is the experimental energy of reaction.

The procedure outlined above (eqs 9–12) is identical to that used in the parameter “calibration” of empirical valence bond models, such as that in refs 32 and 41, or more generally, of the semiempirical valence bond,^{27–31} which allows the energies (barrier height and reaction energy) to be readily fitted to their targets exactly. In general, however, it is much more challenging to “calibrate” the variation of molecular structure along the entire reaction path, especially the precise geometry of the transition state. The sophistication of the mathematical algorithm used by Schlegel and Sonnenberg is a remarkable reflection of the difficulty in constructing an accurate potential energy surface employing empirical valence bond models.^{45,46} The changes of the structural properties, including bond order and force constant, are critically important if one is interested in computing kinetic isotope effects, particularly the error-sensitive secondary effects (2° KIEs), for enzymatic reactions. Inaccuracy can easily be hidden in the large primary KIEs because they typically involve a significant loss of zero-point effects. Thus, agreement with the experiment in primary KIEs, which could be simply due to the loss of the reactant state stretching mode, is not necessarily an indication of good geometry at the transition state. In fact, it is essential to examine both the optimized structure and energy at the transition state to validate the quality of a two-state (or multistate) model against high-level electronic structural data.^{20,26,45,46}

To this end, the off-diagonal matrix element in EH-MOVb (eq 9) is an explicit function of all degrees of freedom of the system, i.e., $H_{rp}(\mathbf{R}) = \langle \Psi_r(\mathbf{R}) | H | \Psi_p(\mathbf{R}) \rangle$.^{10–13,20,26} Consequently, the full-dimensional potential surface can be adequately represented as accurately as the accuracy of the level of the electronic structure method permits, and the transition structure for a reaction can be obtained in accord with that optimized from WFT or DFT calculations. Note that the approach outlined in eqs 9–12 is in principle analogous to that used in effective Hamiltonian valence bond methods to parameterically model the *ab initio* matrix elements to reproduce the exact high-level results.^{33–40,42–49}

3. Computational Details

All calculations are carried out using CHARMM c34a2,⁶⁷ modified with the implementation of the present EH-MOVb. The current QM/MM module in CHARMM at the semiempirical level, called SQUANTUM,⁶⁸ was implemented in our

group by Nam and Walker in 2004, based on a Fortran90 code.⁶⁹ SQUANTUM has been incorporated into the standard distribution and has become the default QM/MM module of CHARMM since version c33a1. The EH-MOVB method was implemented by Song in collaboration with Xie, and it has become a part of the SQUANTUM module with additional options to define the number of states and the number of blocks in each state as well as the associated options. The EH-MOVB method can provide a rigorous valence bond-like model for studying chemical reactions such that the users can conveniently calibrate the model to yield a potential energy surface with the desired barrier height and reaction energy as well as optimized geometry at the transition state. It should be noted that the present EH-MOVB is not a simple quantum mechanical representation of the ideas of empirical valence bond or semiempirical valence bond models such as the London-Eyring-Polanyi-Sato formalism. EH-MOVB is deeply rooted in the traditional approach of Heitler-London-Slater-Pauling function of valence bond theory.

The EH-MOVB module at the semiempirical level is computationally fast; for large systems, the computational bottleneck using our QM/MM potential is in the treatment of the classical long-range electrostatic effects with particle-mesh Ewald (PME) rather than the QM calculation itself. In addition, two options are available for determining the diabatic and adiabatic ground state energies: (1) the consistent diabatic state (CDC) method and (2) the variational diabatic state (VDC) model.²⁰ For those who are interested in using the energy gap between the product and reactant diabatic state as the reaction coordinate,⁷⁰ the VDC diabatic states should be used, since the variational diabatic state is of interest in this case.^{10–12} The VDC determinants also provides the basis states in configuration interaction calculations to give the adiabatic ground state potential energy surface. On the other hand, if geometrical parameters are used to define the reaction coordinate on the adiabatic ground state potential surface, the CDC model is appropriate since this method yields the optimal adiabatic ground-state energy, and analytical gradients can be computed. Note that the CDC method is analogous to multiconfiguration self-consistent field (MCSCF) theory,^{20,26} whereas the VDC approach is akin to a configuration interaction (CI) method.^{10–12}

DFT calculations are carried out using Gaussian 03⁷¹ modified to include the M06-2X functionals.^{72,73} The 6-31+G(d) basis set is used throughout for all calculations. Geometries for the hydride transfer reaction between trimethylamine and flavin cofactor along the reaction coordinate defined below are optimized using the 6-31+G(d) basis set at each level of theory. The recently developed M06-2X functional, which produces similar energies in comparison with MP2 single point calculations, is used to calibrate the EH-MOVB model.

To describe the change in energy and wave function of the two Lewis bond states as the reaction takes place, we define the reaction coordinate here as the difference between the bond lengths of the central hydrogen atom, which is transferred, to the donor atom (C) of **TMA-H** and to the acceptor atom (N) of **Nf⁺**:

$$R_c = R(C - H) - R(H - N) \quad (13)$$

Of course, one can use other definitions to monitor the progress of the reaction, including the difference between the corresponding bond orders or energies of the two Lewis bond states. The geometrical variable, corresponding to the asymmetric bond stretch coordinate, is a good choice and chemically intuitive.

4. Results and Discussion

The main goal of this study is to develop an effective Hamiltonian within MOVB theory to study chemical reactions in solution and in enzymes using CHARMM as a combined QM/MM potential. We hope to illustrate that the procedure can be conveniently used by biochemists as a research tool to help interpret experimental findings, with a straightforward calibration of the EH-MOVB model. We use the hydride transfer reaction from trimethylamine to a flavin cofactor model. The discussion of dynamics simulations is beyond the scope of this report and will be reported separately. We first carry out *ab initio* electronic structural calculations using DFT to yield the structures and energies along the hydride transfer reaction pathway. Then, we optimize the EH-MOVB Hamiltonian to reproduce the “high-level” data. The qualitative features and quantitative results of the diabatic configurations and the adiabatic potential surface will be discussed.

The adiabatic ground state potential energy surfaces determined using DFT with the B3LYP and M06-2X functionals are compared with the standard semiempirical AM1 model and the EH-MOVB method in Figure 1 as a function of the reaction coordinate R_c (eq 13) for the hydride transfer reaction between trimethylamine and a flavin cofactor. Optimized structures at the reactant state and product state complex and the transition state are illustrated in Figure 2 along with key structural parameters. The M06-2X density functional calculations yield an estimated barrier height of 17.4 kcal/mol and a relative energy of -6.1 kcal/mol between the product and reactant states. The popular hybrid B3LYP method underestimates the hydride transfer barrier at 15.1 kcal/mol. The semiempirical AM1 energy profile is qualitatively correct, but it contains two main problems; the computed energy of activation is 22.6 kcal/mol, about 5 kcal/mol too high, compared with the M06-2X value, and the predicted energy of reaction is too endothermic by 6.4 kcal/mol. The latter error is completely transferred into the MOVB relative energies of the reactant and product diabatic states, which can be easily corrected by shifting the product state up by an equal amount, which has no effect on gradient evaluations (Table 1). With an increase in the strength of diabatic coupling between the reactant and product states at the transition state, the barrier height can be lowered, and using the parameters listed in Table 1, we obtained an activation energy of 18.1 kcal/mol for the hydride transfer between trimethylamine and flavin and an energy of reaction of -4.4 kcal/mol. We note that the AM1 model finds another configuration in which the donor N–C–H unit is roughly coplanar with the flavin ring, and it is slightly lower in energy (by about 2 kcal/mol) than the configuration in which **TMA**

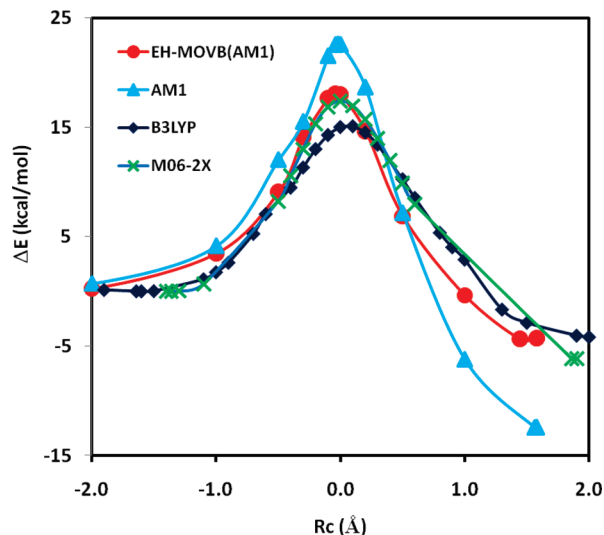


Figure 1. Computed potential energy profile along the minimum energy path ($R_c = R[C-H] - R[H-N]$) for the hydride transfer reaction between trimethylamine and the flavin model using EH-MOVB(AM1) (in red), AM1 (in light blue), B3LYP/6-31G(d) (in navy blue), and M06-2X/6-31G(d) (in green).

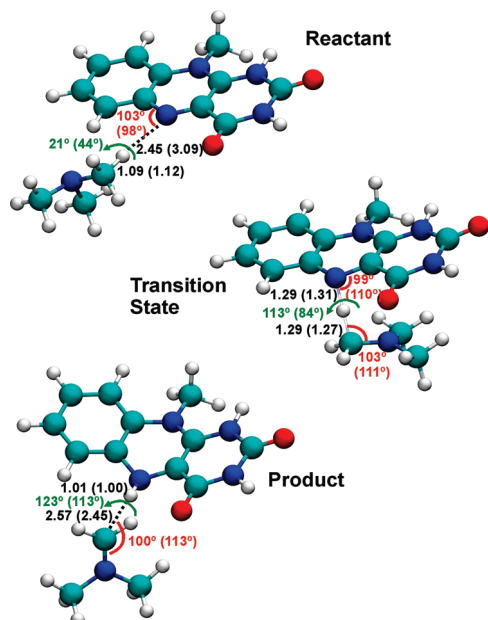


Figure 2. Optimized geometries for the reactant and product complexes and the transition state for the hydride transfer reaction depicted in Scheme 1. MOVb results are listed first, followed by DFT values in parentheses. Distances are given in angstroms and angles in degrees.

Table 1. EH-MOVB Parameters Used in This Study^a

α_{CH}	α_{HN}	$\Delta\epsilon$ (kcal/mol)	γ_{TP} (eV)
0.9	1.0	8.0	1.5

^a The AM1 model is used to define the diabatic reactant and product states for the hydride transfer reaction between trimethylamine and a model flavin cofactor.

is under the plane of the flavin ring. The latter configuration is more closely aligned with the structure found in the active site in the human histone lysine-specific demethylase (LSD1) structure,⁷⁴ which is most relevant to the hydride transfer reaction pathway.

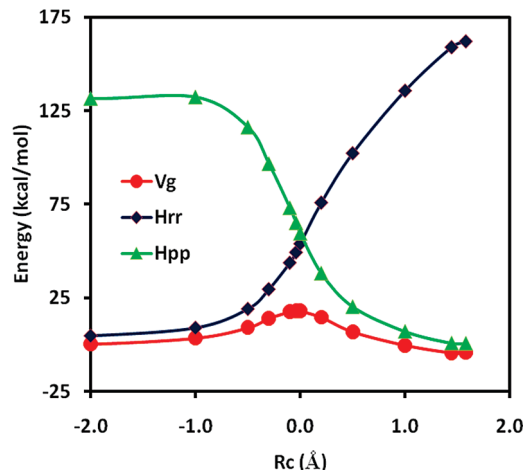


Figure 3. Computed potential energy surfaces for the diabatic reactant state (blue), the diabatic product state (green), and the adiabatic ground state (red) along the minimum energy path.

The optimized geometries at the reactant, product, and transition states from EH-MOVB(AM1) calculations are in accord with those obtained using M06-2X density functional theory. In particular, the donor (C-H) and acceptor (H-N) distances from the hydride atom transferred are 1.29 and 1.29 Å, respectively, which may be compared with the DFT (B3LYP) values of 1.27 and 1.31 Å. The potential energy surface about bond angles and torsional angles is relatively flat, and the accord between EH-MOVB(AM1) and M06-2X is reasonable (Figure 2).

The minimum energy path (MEP) for the hydride transfer from trimethylamine to flavin has been optimized as a function of the reaction coordinate defined by eq 13. In the present study, we have constrained the hydride migration to be collinear with the donor (C) and acceptor (N) atoms, whereas all other degrees of freedom are fully minimized using the ABNR algorithm in CHARMM.⁶⁴ The potential energy curves for the reactant and product diabatic states are shown in Figure 3 along with that for the adiabatic ground state. The reactant state potential shows a steady increasing as the reaction coordinate changes from the reactant to the product side. On the other hand, the potential energy surface is somewhat leveled off for the product state when the molecular geometry is in the reactant state configuration. The trend of the two diabatic potential energy curves is consistent with heterolytic bond cleavages of the reactant (C-H) and the product (N-H) species. At the diabatic state crossing point, which corresponds roughly to the location of the transition state of the hydride transfer reaction, the diabatic state is ca. 40 kcal/mol in energy above the adiabatic ground state, suggesting that there is significant electronic coupling between the reactant and product states. The coupling energy is similar to values determined for proton transfer and nucleophilic substitution reactions using *ab initio* WFT and DFT.^{10–13,20,26}

Figure 4 exhibits the same potential curves shown in Figure 3, but they are plotted against the diabatic energy difference, or the energy-gap reaction coordinate.

$$\Delta E = H_{rr} - H_{pp} \quad (14)$$

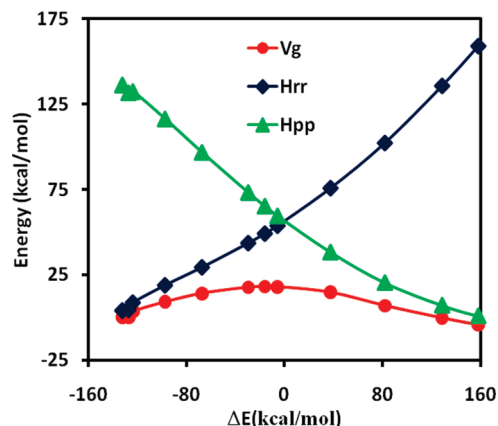


Figure 4. Computed potential energy surfaces in Figure 3 for the diabatic reactant state (blue), the diabatic product state (green), and the adiabatic ground state (red) represented as a function of the energy difference between the reactant and product diabatic states (i.e., the energy-gap reaction coordinate).

Figure 4 shows that the minimum energy potential surface for the adiabatic ground state and those for the diabatic states can be fully represented with the use of either a geometrical or an energy-gap reaction coordinate when the reaction profile is determined by optimizing the geometrical reaction coordinate.

For reactions in solutions or in enzymes, it is of interest to consider the effects of solvent or protein reorganization, and this is often presented using the energy-gap reaction coordinate (eq 14). Although this is easily modeled using an empirical force field to represent the diabatic states, it is far from straightforward if a combined QM/MM potential is employed. The MOVb theory is the first and only QM/MM approach at this time to provide well-defined diabatic states for condensed phase simulations, and *ab initio* MOVb-QM/MM methods have been utilized in the study of solvent effects and reorganization energies for several reactions in solution.^{10–13} Of course, empirical potential functions have been used extensively to describe the energy-gap coordinate.^{41,75,76} The present EH-MOVb approach in the context of a QM electronic structure theory can be conveniently calibrated to yield accurate results and applied to enzymatic catalysis using the program CHARMM. The free energy reaction profile as a function of the energy-gap reaction coordinate is typically obtained through a coupled free energy perturbation simulation,^{10,11,41} which drives the solvent and protein configurations from the reactant state to the product state using a reference potential (which is also called a mapping potential),⁴¹ $V_{RP}(\mathbf{R})$, and umbrella sampling that transforms the biased simulations with $V_{RP}(\mathbf{R})$ into the true adiabatic ground-state potential surface, $V_g(\mathbf{R})$.

The reference potential is typically expressed as a mixture of the diabatic reactant and product energy through a coupling parameter λ :

$$V_{RP}(\lambda) = (1 - \lambda)H_{rr}(\mathbf{R}) + \lambda H_{pp}(\mathbf{R}) \quad (15)$$

where λ is a parameter that varies from 0 (reactant) to 1 (product), and \mathbf{R} specifies the instantaneous geometry of the system. In the present study of the model hydride transfer

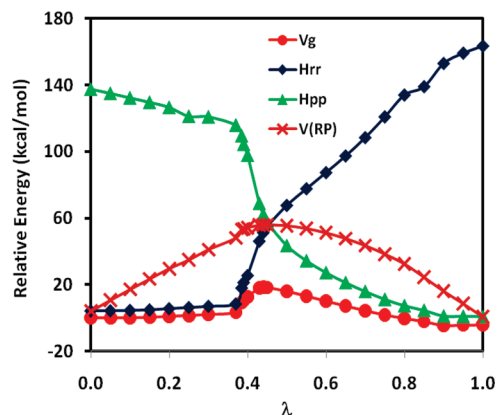


Figure 5. Computed potential energy profiles for the reactant (blue) and product (green) diabatic states along with the adiabatic ground state (red) and the reference potential as a function of the coupling parameter linearly connecting the reactant and product potentials. This reaction path is called the reference minimum energy path, which has a different meaning from that of Figure 1.

from **TMA-H** to flavin (Nf^+), we optimized the reference minimum-energy path (RMEP) defined by eq 15. Then, using the geometries along this reference minimum-energy path, we carried out single-point energy calculations to determine the adiabatic ground state energy. Note that this “RMEP” is not the true adiabatic ground-state MEP (Figures 3 and 4) determined using the EH-MOVb potential, $V_g(\mathbf{R})$, because the structures are optimized using different potential energy surfaces.

Figure 5 depicts the diabatic potential energies and the adiabatic ground state energy, along with the reference potential (eq 15), as a function of the coupling parameter. Since the reference potential is dominantly determined by the reactant diabatic state when λ is less than 0.5, there is a rapid geometry change in the hydride transfer coordinate, which is not explicitly specified by the coupling parameter λ and cannot be effectively restrained to yield a smooth variation. Consequently, there is a sudden change in the molecular geometry as the hydride is fully transferred to the carbon atom, corresponding to a geometrical description of $R_c = -0.7 \text{ \AA}$ to $R_c = -1.8 \text{ \AA}$. The ground-state potential is shown as a function of the geometrical reaction coordinate in Figure 6. This is accompanied by a rather steep increase in the reactant diabatic state and the adiabatic ground state potential in the region of $\lambda = 0.4$ and 0.5 (Figure 5). Interestingly, the overall reference potential shows smoother variations (curve in maroon) due to the compensating contributions from the product diabatic state. The computed barrier height is 18.3 kcal/mol along the RMEP, similar to that of the MEP for the hydride transfer.

Figure 7 recasts the data illustrated in Figure 5, but the adiabatic ground state potential energy surface $V_g(\mathbf{R})$ is plotted against the energy-gap reaction coordinate ΔE . In contrast to Figures 5 and 6, the potential $V_g(\mathbf{R})$ appears to be surprisingly smooth, despite the fact that part of the geometrical variations along the reaction path in fact is discontinuous in Figure 5. Figure 7 shows that a nonsmooth geometrical transition that gives rise to an abrupt energy change can be hidden behind the seeming smooth energy

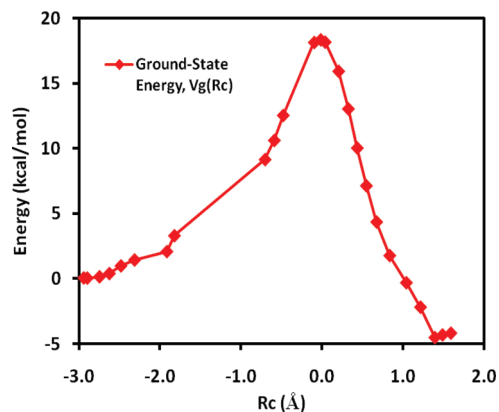


Figure 6. Potential energy profile for the hydride transfer reaction between methylamine and the model flavin cofactor plotted against the geometrical reaction coordinate (eq 14) following the reference minimum energy path in Figure 5.

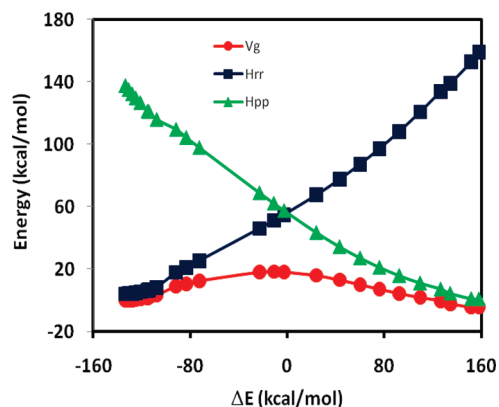


Figure 7. Computed potential energy profiles for the reactant (blue) and product (green) diabatic states and the adiabatic ground state (red) as a function of the energy gap–reaction coordinate for structures obtained along the reference minimum energy path in Figure 5.

curve when the adiabatic ground-state potential is given as a function of a geometry-implicit coordinate such as the energy-gap representation. This observation suggests that it is critically important to report and show both energy results and the corresponding geometries of the reactive molecule in calculations that employ the reference potential of eq 15.

Before we leave this section, we consider the procedure used in condensed phase and enzyme calculations.^{10,11,41} In this case, the reference potential of eq 15 will be used in a series of discrete free energy perturbation simulations with fixed values of λ_i to yield the free energy differences as λ_i changes from 0 to 1. Thus, the free energy at λ relative to the reactant state ($\lambda_0 = 0$) is determined as follows:

$$\Delta G_{\text{RP}}(\lambda) = -RT \sum_{i=0}^{\lambda} \ln \langle e^{-[V_{\text{RP}}(\lambda_{i+1}) - V_{\text{RP}}(\lambda_i)]/RT} \rangle_i \quad (16)$$

where $\langle \dots \rangle_i$ specifies an ensemble average over the potential $V_{\text{RP}}(\lambda_i)$; the summation runs to a value $\lambda = \lambda_{i+1}$. Here, the use of the arbitrary reference potential is purely for the purpose of moving the system to go from molecular configurations corresponding to the reactant state ensemble into the product state. To obtain the free energy of the true

ground state potential surface, governed by the distribution $e^{-V_{\text{g}}(\Delta E)/RT}$, an umbrella sampling-like procedure is applied to the configurations sampled on the basis of the distribution of $e^{-V_{\text{RP}}(\lambda_i)/RT}$. Thus,

$$\Delta G(\Delta E) = \Delta G_{\text{RP}}(\lambda_i) - RT \ln \{ \rho_{\text{RP}}^i(\Delta E) \langle e^{-[V_{\text{g}} - V_{\text{RP}}(\lambda_i)]/RT} \rangle_i \} \quad (17)$$

where the quantity $\rho_{\text{RP}}^i(\Delta E)$ is the normalized distribution of configurations that have a value of ΔE in the ensemble sampled by the reference potential $V_{\text{RP}}(\lambda_i)$.

An important distinction that should be made is that the procedure outlined in eqs 16 and 17 yields the *free energy* profile, or the potential of mean force, as a function of an ensemble of configurations, all having the energy gap ΔE . Obviously, it is not and should not be compared with the *potential energy* surface. Furthermore, the “reaction path” mapped by eq 16 is not the minimum energy path of the adiabatic ground state, nor the reference minimum energy path. Thus, the energy computed, either by averaging over all configurations sampled on the basis of eq 16 or by selecting a single structure of its ensemble, is not directly comparable to results rigorously defined by the MEP. Obviously, it can be deceptive when potential energies free energies obtained along the minimum energy path and or single-point energy calculations on selected geometries from a statistical ensemble are mixed together and compared without rigorously specifying their origins.

5. Conclusions

The effective Hamiltonian–molecular orbital and valence bond (EH-MOVb) method based on nonorthogonal block-localized molecular orbitals has been implemented into the program CHARMM for molecular dynamics simulations of chemical and enzymatic reactions, making use of semiempirical quantum mechanical methods. Building upon previous results using *ab initio* MOVb theory, we introduce two parameters in the EH-MOVb method, along with the addition of the two-center, one-electron integrals across different molecular blocks which may be considered as parameters, such that the barrier height and the relative energy between the reactant and product state for a given chemical reaction can be fitted in good agreement with experimental or high-level *ab initio* and DFT results. The EH-MOVb method provides a highly accurate and computationally efficient QM/MM model for dynamics simulation of chemical reactions in solution. The MOVb theory is the first and currently the only QM/MM method that allows the potential of mean force to be determined as a function of the energy-gap reaction coordinate for characterization of solvent reorganization effects.

The EH-MOVb method is illustrated by examination of the potential energy surface of the hydride transfer reaction from trimethylamine to a flavin cofactor model in the gas phase. In the present study, we employ the semiempirical AM1 model, which yields a qualitatively correct energy profile along the minimum energy path (Figure 1). However, as in most practical applications using semiempirical Hamiltonians, the quantitative results are not satisfactory. Tradi-

tionally, there is no systematic way of improving the semiempirical model, even though the qualitative features of structure and energy are reasonable. In EH-MOVB, the barrier height is optimized to reproduce the desired (accurate) value in the gas phase (i.e., the intrinsic performance of the effective Hamiltonian) either by scaling or by adding a constant to the off-diagonal matrix element. The present EH-MOVB method offers an alternative approach to characterization of solvent and protein-reorganization effects in the realm of truly combined QM/MM simulations.

Acknowledgment. We thank the National Institutes of Health (GM46736) for support of this work. A.P. is a recipient of the Thailand Research Fund, under the Royal Golden Jubilee Ph.D. Graduate Program (PHD/0211/2547).

Supporting Information Available: Structures optimized using the standard AM1 method, the EH-MOVB method along the minimum energy path and along the reference minimum energy path, and the B3LYP/6-31+G(d) method along the minimum energy path. All semiempirical calculations were performed with the SQUANTM module of CHARMM. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198.
- Gao, J.; Xia, X. *Science* **1992**, *258*, 631.
- Gao, J.; Ma, S.; Major, D. T.; Nam, K.; Pu, J.; Truhlar, D. G. *Chem. Rev.* **2006**, *106*, 3188.
- Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1985**, *107*, 154.
- Gao, J. *J. Am. Chem. Soc.* **1991**, *113*, 7796.
- Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *233*, 231.
- Marti, S.; Moliner, V.; Tunon, I. *J. Chem. Theory Comput.* **2005**, *1*, 1008.
- Garcia-Viloca, M.; Truhlar, D. G.; Gao, J. *Biochemistry* **2003**, *42*, 13558.
- Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486.
- Mo, Y.; Gao, J. *J. Comput. Chem.* **2000**, *21*, 1458.
- Mo, Y.; Gao, J. *J. Phys. Chem. A* **2000**, *104*, 3012.
- Gao, J.; Garcia-Viloca, M.; Poulsen, T. D.; Mo, Y. *Adv. Phys. Org. Chem.* **2003**, *38*, 161.
- Cembran, A.; Song, L.; Mo, Y.; Gao, J. *J. Chem. Theory Comput.* **2009**, *5*, 2702.
- Mo, Y.; Peyerimhoff, S. D. *J. Chem. Phys.* **1998**, *109*, 1687.
- Mo, Y.; Zhang, Y.; Gao, J. *J. Am. Chem. Soc.* **1999**, *121*, 5737.
- Mo, Y.; Gao, J.; Peyerimhoff, S. D. *J. Chem. Phys.* **2000**, *112*, 5530.
- Stoll, H.; Preuss, H. *Theor. Chem. Acc.* **1977**, *46*, 12.
- Gianinetti, E.; Raimondi, M.; Tornaghi, E. *Int. J. Quantum Chem.* **1996**, *60*, 157.
- Gianinetti, E.; Vandoni, I.; Famulari, A.; Raimondi, M. *Adv. Quantum Chem.* **1998**, *31*, 251.
- Song, L.; Gao, J. *J. Phys. Chem. A* **2008**, *112*, 12925.
- Gao, J.; Mo, Y. *Prog. Theor. Chem. Phys.* **2000**, *5*, 247.
- Cooper, D. L.; Gerratt, J.; Raimondi, M. *Adv. Chem. Phys.* **1987**, *69*, 319.
- Hiberty, P. C.; Flament, J. P.; Noizet, E. *Chem. Phys. Lett.* **1992**, *189*, 259.
- Wu, W.; Song, L.; Cao, Z.; Zhang, Q.; Shaik, S. *J. Phys. Chem. A* **2002**, *106*, 2721.
- Song, L.; Mo, Y.; Zhang, Q.; Wu, W. *J. Comput. Chem.* **2005**, *26*, 514.
- Song, L.; Mo, Y.; Gao, J. *J. Chem. Theory Comput.* **2009**, *5*, 174.
- Sato, S. *J. Chem. Phys.* **1955**, *23*, 592.
- Kuntz, P. J.; Nemeth, E. M.; Polanyi, J. C.; Rosner, S. D.; Young, C. E. *J. Chem. Phys.* **1966**, *44*, 1168.
- Raff, L. M.; Stivers, L.; Porter, R. N.; Thompson, D. L.; Sims, L. H. *J. Chem. Phys.* **1970**, *52*, 3449.
- Raff, L. M. *J. Chem. Phys.* **1974**, *60*, 2220.
- Silver, D. M.; Brown, N. J. *J. Chem. Phys.* **1980**, *72*, 3859.
- Warshel, A.; Weiss, R. M. *J. Am. Chem. Soc.* **1980**, *102*, 6218.
- Sheppard, M. G.; Freed, K. F. *J. Chem. Phys.* **1981**, *75*, 4507.
- Hurtubise, V.; Freed, K. F. *Adv. Chem. Phys.* **1993**, *83*, 465.
- Martin, C. H.; Graham, R. L.; Freed, K. F. *J. Phys. Chem.* **1994**, *98*, 3467.
- Bernardi, F.; Olivucci, M.; Robb, M. A. *J. Am. Chem. Soc.* **1992**, *114*, 1606.
- Bearpark, M. J.; Robb, M. A.; Bernardi, F.; Olivucci, M. *Chem. Phys. Lett.* **1994**, *217*, 513.
- Bearpark, M. J.; Bernardi, F.; Olivucci, M.; Robb, M. A. *J. Phys. Chem. A* **1997**, *101*, 8395.
- Bearpark, M. J.; Smith, B. R.; Bernardi, F.; Olivucci, M.; Robb, M. A. *ACS Symp. Ser.* **1998**, *712*, 148.
- Chang, Y. T.; Miller, W. H. *J. Phys. Chem.* **1990**, *94*, 5884.
- Aqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523.
- Schmitt, U. W.; Voth, G. A. *J. Phys. Chem. B* **1998**, *102*, 5547.
- Day, T. J. F.; Soudackov, A. V.; Cuma, M.; Schmitt, U. W.; Voth, G. A. *J. Chem. Phys.* **2002**, *117*, 5839.
- Maupin, C. M.; Wong, K. F.; Soudackov, A. V.; Kim, S.; Voth, G. A. *J. Phys. Chem. A* **2006**, *110*, 631.
- Schlegel, H. B.; Sonnenberg, J. L. *J. Chem. Theory Comput.* **2006**, *2*, 905.
- Sonnenberg, J. L.; Schlegel, H. B. *Mol. Phys.* **2007**, *105*, 2719.
- Kim, Y.; Corchado, J. C.; Villa, J.; Xing, J.; Truhlar, D. G. *J. Chem. Phys.* **2000**, *112*, 2718.
- Tishchenko, O.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13530.
- Lin, H.; Zhao, Y.; Tishchenko, O.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 1237.
- Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.

- (52) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101.
- (53) Stoll, H.; Wagenblast, G.; Preuss, H. *Theor. Chim. Acta* **1980**, *57*, 169.
- (54) Raimondi, M.; Famulari, A.; Specchio, R.; Sironi, M.; Moroni, F.; Gianinetti, E. *THEOCHEM* **2001**, *573*, 25.
- (55) Mo, Y.; Gao, J. *J. Phys. Chem. A* **2001**, *105*, 6530.
- (56) Mo, Y.; Subramanian, G.; Gao, J.; Ferguson, D. M. *J. Am. Chem. Soc.* **2002**, *124*, 4832.
- (57) Mo, Y.; Schleyer, P. v. R.; Wu, W.; Lin, M.; Zhang, Q.; Gao, J. *J. Phys. Chem. A* **2003**, *107*, 10011.
- (58) Mo, Y.; Wu, W.; Song, L.; Lin, M.; Zhang, Q.; Gao, J. *Angew. Chem., Int. Ed.* **2004**, *43*, 1986.
- (59) Mo, Y.; Gao, J. *J. Phys. Chem. B* **2006**, *110*, 2976.
- (60) Khaliullin, R. Z.; Head-Gordon, M.; Bell, A. T. *J. Chem. Phys.* **2006**, *124*, 204105/1.
- (61) Mo, Y.; Gao, J. *Acc. Chem. Res.* **2007**, *40*, 113.
- (62) Pople, J. A.; Santry, D. P.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S129.
- (63) Mo, Y.-r.; Alhambra, C.; Gao, J.-l. *Huaxue Xuebao* **2000**, *58*, 1504.
- (64) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (65) Pople, J. A.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S136.
- (66) Elstner, M.; Porezag, D.; Juugnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Sukai, S.; Seifect, G. *Phys. Rev. B* **1998**, *58*, 7260.
- (67) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodosek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
- (68) Nam, K.; Prat-Resina, X.; Garcia-Viloca, M.; Devi-Kesavan, L. S.; Gao, J. *J. Am. Chem. Soc.* **2004**, *126*, 1369.
- (69) Walker, R. C.; Crowley, M. F.; Case, D. A. *J. Comput. Chem.* **2008**, *29*, 1019.
- (70) Marcus, R. A. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 1111.
- (71) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision D.01; Gaussian, Inc.: Pittsburgh, PA, 2004.
- (72) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 1009.
- (73) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2007**, *111*, 4632.
- (74) Chen, Y.; Yang, Y.; Wang, F.; Wan, K.; Yamane, K.; Zhang, Y.; Lei, M. *Proc. Natl. Acad. Sc. U.S.A.* **2006**, *103*, 13956.
- (75) Billeter, S. R.; Webb, S. P.; Agarwal, P. K.; Iordanov, T.; Hammes-Schiffer, S. *J. Am. Chem. Soc.* **2001**, *123*, 11262.
- (76) Hatcher, E.; Soudackov, A. V.; Hammes-Schiffer, S. *J. Am. Chem. Soc.* **2007**, *129*, 187.

CT1001686